UNIVERSITY OF CAMBRIDGE

Parton Distributions in Beyond the Standard Model Theories

James Michael Moore

St Edmund's College

This thesis is submitted on Wednesday 2nd August, 2023 for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

James Michael Moore
Wednesday 2nd August, 2023

Abstract

Parton Distributions in Beyond the Standard Model Theories

James Michael Moore

Parton distributions are a key ingredient of precise predictions for collider experiments. They are usually determined from fits to experimental data under the assumption that the Standard Model (SM) of particle physics is complete; however, this can bias studies of beyond the Standard Model (BSM) physics if these SM-like PDFs are used in these analyses. It is important to quantify the extent to which this occurs, in order to avoid making incorrect conclusions about BSM physics.

We begin in Chapter 1 with a review of perturbative quantum chromodynamics (QCD) and parton distribution functions (PDFs), providing a definition of the PDFs at next-to-leading order in QCD perturbation theory. At the end of the Chapter, in Sect. 1.4, we introduce the main problem that this thesis aims to address in a variety of special cases, namely the simultaneous extraction of PDFs together with other theory parameters (specifically BSM theories).

In Chapters 2, 3 and 4, we describe the interplay between PDFs and the parameters of various BSM models. In more detail, in Chapter 2, we perform an approximate simultaneous extraction of PDFs together with the parameters of a *dark photon* model; in particular, we use projected high-luminosity LHC (HL-LHC) data to investigate the sensitivity of the HL-LHC to our particular class of light, leptophobic dark photons. Subsequently, in Chapter 3, we introduce the Standard Model Effective Field Theory (SMEFT), and carry out a simultaneous determination of PDFs together with two parameters drawn from the SMEFT; we show that at the HL-LHC, there will be significant interplay between extraction of PDFs and SMEFT parameters. In Chapter 4, we perform a much more comprehensive analysis of the PDF-SMEFT interplay in the top sector, using a new efficient methodology, SIMUnet. Importantly in Sect. 4.7, we also comment on the efficacy of the Monte Carlo replica method for error propagation, which forms the heart of the uncertainty calculation in both the NNPDF and SIMUnet methodologies.

In the second half of this thesis, we focus on future issues in PDF fitting, related to the work presented in the previous chapters. In Chapter 5, we explore how New Physics

in the data might be inadvertently 'fitted away' into the PDFs, if the data is treated as SM-like. We also recommend strategies for disentangling PDFs and BSM effects. Finally, in Chapter 6, we discuss the Monte Carlo replica method used in many of the previous chapters, and discuss the need for its replacement in future PDF and BSM fits.

Acknowledgements

First and foremost, a special thanks goes out to my excellent PhD supervisor, Maria. Not only is she an extremely accomplished researcher, with outstanding, creative ideas for novel work, she is also an absolute pleasure to work with. Her dedication to her students and postdocs is unmatched. Thank you for giving me the opportunity to work with you in such an amazing group.

I am forever indebted to my wonderful parents, Estelle and Garry, and grandparents, Catherine and Michael. Your unwavering support has been a driving force for any success I have achieved, from your love and care when I was younger, to your emotional and financial support in the present. Similarly my siblings, Catherine and Oliver, have always been there for me as firm friends.

Thanks go out to Alastair for everything, now and always; I am extremely glad to still remain close to you.

Thank you to my longstanding friends Yanbo and Sam for providing an excellent distraction from work during my PhD; I look forward to future exchanges of crude 'art', and death marches along English dykes.

To my amazing friends from Downing who moved away to Bristol, Evie, Mick, Stephen and Luke, I would like to thank you for all our time spent together (I know that the only reason you suffer my company is because I am the one with a copy of Agricola, but I'm willing to overlook this). Thank you also to my wonderful Downing friends who now live in Norwich, Beth and Dan, whose hosting skills are unmatched, and whose kindness and generosity is always much appreciated.

Thanks to the fantastic members of my research group, without whom the work in this thesis would have been impossible. Shayan, Cameron and Zahari: thank you for helping me navigate the complicated world of PDF fitting, particularly at the start of my PhD when I felt most lost. Elie and Mark: it has been a pleasure getting to know you both at

the start of your PhD journeys, and I will be glad to continue telling you wrong things about PDF fitting, for another year at least. Manu: sharing an office with you has been tremendous fun, you are a constant source of smiles and laughter, but you are also there to listen whenever I need someone to talk to - thank you for everything you do. Maeve: not only are you an outstanding collaborator, an absolute pleasure to work with, but you are one of the kindest people I have had the privilege to know; I have missed you very dearly in Cambridge, but look forward to a Heidelberg visit in the near future! Luca: whilst it took more time for me to understand your personality, it has been worth it - you have become more than a collaborator, you are now a very good friend.

Thank you to Matthew McCullough and Admir Greljo for being outstanding collaborators; it was a pleasure working with you on the papers featured in this thesis. And further, an additional special thanks to Matthew in particular for writing a reference for me for my postdoc position.

To my PhD cousin, Hannah: you are an incredible physicist, and a constant source of inspiration - don't ever doubt it.

Thanks to those members of the department who are not particle physicists, but nevertheless socialise with me (I try not to hold your poor choice of subjects against you). Wilfred: thank you for being so consistently warm, friendly and up for hanging out; when I started supervising, I didn't envisage becoming such good friends with one of my students, but I am extremely glad of it, and I will miss you very much when you leave Cambridge. Mitchell: thank you also for always enthusiastically accepting my invites, it's always great fun seeing you - I'm looking forward to another fun year of board games, musicals and *Doctor Who*. Campbell: I'd like to thank you for the great privilege of participating in your podcast, *Question Field*; I hope that the ratings have not gone down as a result! Miren: thank you for being such an outstanding dinner guest, and more recently a patient squash instructor! Melo: thanks for the fun times in Cambridge, especially in my last year, featuring lots of board games and *Doctor Who* - looking forward to visiting you soon.

Thank you to Ben Allanach and Matt Wingate for giving me the opportunity to teach your respective Symmetries courses during my PhD. And a special thanks to Ben for letting me give a one-off substitute lecture, it was a lot of fun! And thank you to Manda, not only for everything you have done behind the scenes during my stay at DAMTP, but for being a pleasure to talk to, and always taking an active interest in my life around the PhD.

Finally, a huge thanks to my other friends from out and around in Cambridge. Philip: thank you so much for the lovely time we spent together - it was a highlight of my PhD years, and really meant a lot to me. Victoria: thank you for the all the times you've put up with me venting over work, usually whilst drinking something appropriate. Will and Elena: it's always an absolute pleasure seeing you, you both bring an infectious enthusiasm for anything and everything we do - and to Elena specifically, thank you for the beautiful chocolate cake you made me for my 26th birthday, it was such a wonderfully kind gesture! Carolina: thank you for putting up with me as a salsa partner for so long; I always had a wonderful time dancing with you at Darwin.

I love all of you dearly; the positive impact you have had on my life is immeasurable.

Contents

Chapter 1

Introduction: perturbative QCD and parton distributions

> To learn which questions are unanswerable, and not to answer them: this skill is most needful in times of stress and darkness.
>
> *from The Left Hand of Darkness, by Ursula K. Le Guin*

The Standard Model (SM) is currently the most successful description of particle physics, with its predictions compatible, to within five standard deviations, with all experimental data to date (see for example, the ATLAS and CMS data-theory summary plots contained in Ref. [1] and [2]). It can be concisely described as a Poincaré-invariant $SU(3) \times SU(2) \times U(1)$ gauge theory, with a specific matter content (consisting of six flavours of *quarks*, transforming in the fundamental representation of $SU(3)$, and six flavours of *leptons*, transforming in the trivial representation of $SU(3)$), together with a scalar boson called the *Higgs boson*. The acquisition of a vacuum expectation value by the Higgs boson induces the spontaneous breaking of the $SU(2) \times U(1)$ subgroup to a $U(1)$ symmetry, resulting in the familiar theory of *quantum electrodynamics* (QED). The subgroup $SU(3)$ remains unbroken, and describes the theory of *quantum chromodynamics* (QCD), the subject of this chapter.

In more detail, the part of the SM Lagrangian density corresponding to QCD is given by:

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu,a} + \sum_q \bar{q}(i\not{D} - m_q)q, \tag{1.1}$$

where $G^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g_S f^a{}_{bc} A^b_\mu A^c_\nu$ is the field-strength tensor for the gluon fields

A^a_μ (here, μ is a Lorentz index, and $a = 1, ..., 8$ is an $SU(3)$ adjoint index, labelling the eight species of gluon), the sum is over the quark fields $q = u, d, s, c, b, t$ (which carry both a Lorentz index and an $SU(3)$ fundamental index), the covariant derivative is defined by:

$$D_\mu = I\partial_\mu - ig_S A^a_\mu T^a, \tag{1.2}$$

with I the identity matrix and T^a the generators of the Lie algebra $\mathfrak{su}_{\mathbb{C}}(3)$, and m_q the mass of the quark species q. We define the *strong coupling* α_S in terms of the coupling constant g_S appearing in the Lagrangian density (1.1) via:

$$\alpha_S := \frac{g_S^2}{4\pi}, \tag{1.3}$$

in analogy with the definition of the fine structure constant of QED.

From the Lagrangian density (1.1), one should in principle be able to directly predict all observable QCD phenomena; however, in practice, this is not (yet) theoretically possible. This can be attributed to two major barriers:

(1) QCD is a *strongly-coupled* theory. At the time of writing, the global best-fit value of the coupling α_S is given (at a renormalisation scale equal to the mass of the Z-boson; see below) by $\alpha_S = 0.1179 \pm 0.0009$, compared with the fine structure constant of electromagnetism α_e which is more than ten times smaller, and the electroweak coupling α_{EW} which is approximately 10^6 times smaller (see Sections 1 and 9.4 of Ref. [3]). Thus, the application of perturbation theory in QCD is in question; the series expansions which form the backbone of all order-by-order calculations in QFT predictions are on the borderline of convergence.[1]

(2) The asymptotic states in QCD are *bound states* called *hadrons*, instead of free quarks and gluons. To obtain predictions in standard quantum field-theoretic perturbation theory, one perturbs around the *free* theory; in particular, working perturbatively one must assume that incoming and outgoing states are free quark and gluon states. Naturally, this is a poor approximation in the case of observable QCD processes, and we must come up with something more robust.

Overcoming these difficulties in order to make predictions for collider experiments is the industry of *perturbative QCD*. The primary pillars of the field are:

[1]Technically, since we expect these series expansions to be *asymptotic*, we do not expect them to converge - for an asymptotic series, we can merely hope to have a number of terms which give a good approximation before we must truncate. For QCD, the fact that the coupling is large means that the number of terms before we must truncate is likely to be smaller than those of electromagnetism or the electroweak theory.

(1) **Asymptotic freedom.** As for most[2] of the parameters of the SM, the strong coupling $\alpha_S(\mu_R)$ is defined to be a *running coupling*, using the modified minimal subtraction renormalisation scheme ($\overline{\text{MS}}$). The coupling depends on an arbitrary scale μ_R called the *renormalisation scale*; this scale is arbitrary in the sense that if we were to perform calculations to all orders, observables would carry no dependence on this scale. *However*, truncating the perturbation series early can result in a superficial dependence on the scale.

In many cases,[3] dimensional analysis dictates that the arbitrary scale μ_R will appear in logarithms of the form $\log(\mu_R/Q)$ at any given order in perturbation theory, where Q is a characteristic energy scale of the process. Therefore, to avoid the presence of large logarithms in perturbation theory the renormalisation scale is usually taken as $\mu_R = Q$. This can cause problems for the convergence of perturbation theory if the value of Q is such that $\alpha_S(Q)$ is large (so we can be faced with the problem of choosing μ_R to either cancel large logarithms, or keep α_S sufficiently small).

Fortunately, QCD possesses the property that $\alpha_S(Q)$ decreases as Q increases; this property, called *asymptotic freedom*, was first shown in the Nobel prize-winning calculations of Politzer, Gross and Wilczek [5, 6]. The evolution of $\alpha_S(\mu_R)$ with scale is given by:

$$\frac{d\alpha_S(\mu_R)}{d\log\mu_R} = -\left(11 - \frac{2n_f}{3}\right)\frac{\alpha_S(\mu_R)^2}{2\pi} + O(\alpha_S(\mu_R)^3), \tag{1.4}$$

where n_f is the number of active quark flavours. Therefore, provided we work at sufficiently high energies, the strongly-coupled nature of QCD can be overcome.

Indeed, asymptotic freedom alone is enough to perform complete calculations in special cases where we sum over all possible hadronic states. A classic example is electron-positron annihilation into hadrons, $e^+e^- \to$ any hadrons; naïvely, this process seems inaccessible since the final states are not free quarks and gluons, but are hadronic states instead. However, if we do not care about *which* hadrons we produce, in our calculation we may at some point apply the *completeness relation*:

$$\sum_{\substack{X,\text{ a hadronic}\\ \text{state}}} |X\rangle\langle X| = \sum_{\substack{Y,\text{ a free quark}\\ \text{and gluon state}}} |Y\rangle\langle Y|\,. \tag{1.5}$$

That is, instead of working with a basis of hadronic states for our outgoing state space, the fact that we are summing over all possible hadronic final states allows us to replace

[2]With the notable exceptions of the masses of the Higgs, and heavy bosons, where typically an on-shell mass renormalisation is used.

[3]Indeed, it can be shown that this behaviour is generic; see the discussion preceding and following Eq. (31) in [4], for example.

this basis with a basis of free quark and gluon states (taking $|Y\rangle = |q\rangle, |g\rangle, |qg\rangle$, etc.). At the level of cross-sections, we explicitly have:

$$\sigma(e^+e^- \to \text{any hadrons}) = \sum_{\substack{X, \text{ a hadronic} \\ \text{state}}} \sigma(e^+e^- \to X) = \sum_{\substack{Y, \text{ a free quark} \\ \text{and gluon state}}} \sigma(e^+e^- \to Y). \tag{1.6}$$

Asymptotic freedom now permits us to apply perturbation theory to the cross-sections on the right hand side, provided that we work at sufficiently high energies.

(2) **Factorisation theorems.** Perturbative QCD would be a particularly uninteresting field if the only quantities we could calculate were those in which we summed over all possible hadronic states. Fortunately, another tool exists to help us deal with cases where we must confront the unknown hadronic states, namely *factorisation theorems*; these theorems provide a separation of a process with identified hadrons in either the initial or final states (or indeed both) into a *perturbatively calculable*, but *process-dependent* part, and a *non-perturbative*, but *process-independent* (often called *universal*) part, which itself depends only on the identified hadrons.

In more detail (but still sketching a schematic picture for now), the perturbatively calculable, process-dependent part is called the *hard cross-section* or *partonic cross-section* and is often written as $\hat{\sigma}$. There are several non-perturbative, universal pieces: one for each of the identified hadrons involved in the process. For identified hadrons in the initial state, these non-perturbative objects are called *parton distribution functions* (PDFs), often written as f, whilst for hadrons in the final state, these non-perturbative objects are called *fragmentation functions*, often written as D. Overall, factorisation theorems tell us that for a process with incoming hadrons $h_1, ..., h_n$ and outgoing hadrons $H_1, ..., H_m$, the cross-section can be decomposed into a convolution of the form:

$$\sigma = \hat{\sigma} \otimes f_{h_1} \otimes ... \otimes f_{h_n} \otimes D_{H_1} \otimes ... \otimes D_{H_m}. \tag{1.7}$$

The symbol $\otimes$ denotes the *Mellin convolution*, which we shall define later in the text.

In the rest of this chapter we will focus exclusively on the second tool, namely factorisation theorems. We begin in Section 1.1 with a discussion of a factorisation theorem for a specific process, namely *deep inelastic scattering* (DIS); this will provide us with a working definition of PDFs, which shall be the key players in the rest of this text. In Section 1.2, we shall discuss salient properties of the PDFs, namely their dependence on factorisation scale (governed by the *DGLAP evolution equations*), and their dependence on momentum fraction (constrained by *sum rules*). In Section 1.3, we shall describe how

PDFs can be obtained via fits to the global experimental dataset. Finally, in Section 1.4, we will introduce the key problem that the remaining chapters of this thesis will tackle, namely the joint determination of PDFs together with theory parameters, such as coupling constants and masses.

1.1 Factorisation theorems

As alluded to in the introduction to this chapter, factorisation theorems provide us a way of separating a given process into a perturbative, process-dependent part, and a non-perturbative, universal part. To introduce the ideas, we shall focus on the important special case of *deep inelastic scattering* (DIS); towards the end of this section, we shall discuss how the results generalise.

We begin by introducing *structure functions* for DIS, which are the experimentally reported observables for the process. Predictions for the DIS structure functions can be written as the contraction of two tensors: the *leptonic* tensor, which is perturbatively calculable, and the *hadronic* tensor, which is not. We proceed to parametrise the hadronic tensor in terms of Feynman's phenomenological *parton model* [7], hence introducing the central objects of study in this thesis, namely parton distributions. We then perform a detailed calculation of the hadronic tensor in the parton model at both leading order in QCD, and at next-to-leading order in QCD; this allows us to introduce the key definition of the $\overline{\text{MS}}$ PDFs (modified minimal subtraction PDFs). Finally, we conclude with some general remarks about techniques that have been used to *prove* factorisation theorems to all orders.

1.1.1 Structure functions for DIS

Consider a lepton ℓ impacting on a hadron H, producing some detected lepton ℓ' (which may or may not be of the same species as the initial lepton ℓ) and any hadronic state X, which we do not detect (see Figure 1.1).

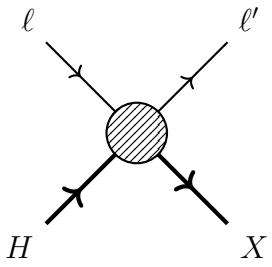


Figure 1.1: Deep inelastic scattering (DIS) of a lepton ℓ on a hadron H, producing a lepton ℓ' and a hadronic state X.

The differential cross-section for this process, in d spacetime dimensions, may be expressed as:

$$d\sigma = \frac{1}{4\sqrt{(p_\ell \cdot p_H)^2 - M_\ell^2 M_H^2}} \frac{d^{d-1}\mathbf{p}_{\ell'}}{(2\pi)^{d-1} 2E_{\ell'}} \sum_X \left(\prod_{i=1}^{n_X} \frac{d^{d-1}\mathbf{p}_{(i,X)}}{(2\pi)^{d-1} 2E_{(i,X)}} \right)$$
$$\cdot (2\pi)^d \delta^d \left(p_\ell + p_H - p_{\ell'} - \sum_{i=1}^{n_X} p_{(i,X)} \right) \overline{|\mathcal{M}(\ell H \to \ell' X)|^2}, \tag{1.8}$$

where the pair (i, X) denotes the ith particle in the hadronic state X (with i in the range $i \in \{1, ..., n_X\}$, for a total of n_X particles in the state X), the four-vector $p_P = (E_P, \mathbf{p}_P)$ is the four-momentum of the particle P, and $\mathcal{M}(\ell H \to \ell' X)$ is the amplitude for the process. The sum over X indicates a sum over all possible hadronic states which can be produced in the collision, reflecting our desire to be blind towards the hadronic products. The bar over the modulus-squared amplitude denotes the spin/colour-sum/average; more precisely, we average over the possible spin/colour states of the initial states, and sum over the spin/colour states of the final states (note that for hadronic states H, X the colour sum/average is trivial because hadronic states are colourless). The notation M_P means the rest mass of the state P.

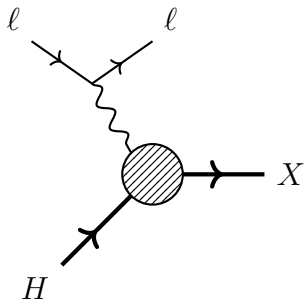


Figure 1.2: Feynman diagram for DIS of an electron off a proton, mediated by a photon, to leading order in QED.

For ease of exposition, let us now focus on the specific case where ℓ, ℓ' are electrons, H is a proton, and the process is mediated by a photon. The Feynman diagram for this process (working to leading order in QED for the photon-electron interactions) takes the form shown in Fig. 1.2. Thus, the amplitude may be expressed in the form:

$$\mathcal{M}(\ell H \to \ell' X) = (ie)^2 \bar{u}(p_{\ell'}) \gamma^\mu u(p_\ell) \cdot \frac{i}{q^2} \cdot \langle X | J_\mu | H \rangle \,, \tag{1.9}$$

where the factors of ie arises from the electron-photon and electron-hadron interactions, the spinor algebra $\bar{u}(p_{\ell'})\gamma^\mu u(p_\ell)$ arises from the electron lines, the factor of i/q^2 comes from the photon propagator (where we have defined $q = p_\ell - p_{\ell'}$ to be the momentum of the virtual photon), and the matrix element $\langle X | J_\mu | H \rangle$ comes from the hadron-photon

interaction in the lower half of the diagram. Here, J_μ is the electromagnetic current governing photon-quark interactions, given by:

$$J_\mu = \sum_q e_q \bar{q} \gamma_\mu q, \tag{1.10}$$

where e_q is the charge on a quark of flavour q, in units of the electron charge. Taking the spin/colour-sum/average of the modulus-squared amplitude, we obtain:

$$\overline{|\mathcal{M}(\ell H \to \ell' X)|^2} = \frac{e^4}{2q^4} \mathrm{Tr}\left(\gamma^\mu \not{p}_\ell \gamma^\nu \not{p}_{\ell'}\right) \overline{\langle H|J_\nu^\dagger|X\rangle \langle X|J_\mu|H\rangle}, \tag{1.11}$$

where the bar over the amplitude denotes the spin/colour-sum/average. At this point, it is convenient to introduce two tensors in terms of which the differential cross-section can be parametrised. We define the *leptonic tensor* via:

$$L^{\mu\nu} = e^4 \mathrm{Tr}\left(\gamma^\mu \not{p}_\ell \gamma^\nu \not{p}_{\ell'}\right) = 4e^4 \left(p_\ell^\mu p_{\ell'}^\nu + p_\ell^\nu p_{\ell'}^\mu - \eta^{\mu\nu} p_\ell \cdot p_{\ell'}\right), \tag{1.12}$$

where the second form is obtained through basic gamma matrix identities, and the *hadronic tensor* via:

$$H_{\mu\nu} = \frac{1}{4\pi} \sum_X \left(\prod_{i=1}^{n_X} \frac{d^{d-1}\mathbf{p}_{(i,X)}}{(2\pi)^{d-1} 2E_{(i,X)}} \right) \cdot (2\pi)^d \delta^d \left(p_\ell + p_H - p_{\ell'} - \sum_{i=1}^{n_X} p_{(i,X)} \right) \overline{\langle H|J_\nu^\dagger|X\rangle \langle X|J_\mu|H\rangle}. \tag{1.13}$$

The differential cross-section may then be expressed in the simplified form:

$$d\sigma = \frac{1}{4\sqrt{(p_\ell \cdot p_H)^2 - M_\ell^2 M_H^2}} \cdot \frac{1}{q^4} \frac{d^{d-1}\mathbf{p}_{\ell'}}{(2\pi)^{d-2} 2E_{\ell'}} L^{\mu\nu} H_{\mu\nu}. \tag{1.14}$$

This form is particularly useful, since it clearly decomposes the differential cross section's dependence into a *perturbative* part, namely the leptonic tensor, and a *non-perturbative* part, namely the hadronic tensor. Hence, we have isolated the key challenge in computing a DIS cross-section: modelling the hadronic tensor.

Before commenting further on the hadronic tensor though, it is useful to perform some further superficial work to make Eq. (1.14) match with standard expressions in the literature (see for example Chapter 19 of Ref. [3]). We begin by introducing the standard Lorentz-invariant kinematical variables Q^2, x and y, defined by:

$$Q^2 = -q^2, \qquad x = -\frac{q^2}{2q \cdot p_H}, \qquad y = \frac{q \cdot p_H}{p_\ell \cdot p_H}. \tag{1.15}$$

The first invariant quantity is simply the *virtuality* of the mediating photon; it has

units of squared energy. The second invariant quantity is called the *Bjorken-x*, and its interpretation will become clear when we discuss the parton model shortly. The third invariant quantity is called the *inelasticity*, and measures the fraction of energy lost by the electron; we can see this by working in the rest frame of the hadron, $\mathbf{p}_H = \mathbf{0}$, which yields:

$$y = \frac{E_\ell - E_{\ell'}}{E_\ell} = 1 - \frac{E_{\ell'}}{E_\ell}, \tag{1.16}$$

where E_ℓ is the energy of the initial electron, and $E_{\ell'}$ is the energy of the final electron.

Specialising to the case where $d = 4$, we can change variables from $\mathbf{p}_{\ell'}$ in Eq. (1.14) to the variables x, y, ϕ yielding the simplified form:

$$\frac{d^3\sigma}{dxdyd\phi} = \frac{y}{32\pi^2 Q^4} L^{\mu\nu} H_{\mu\nu}. \tag{1.17}$$

The details of this transformation are given in App. A.

To make further progress, we study the Lorentz structure of the hadronic tensor. As can be observed from the definition in Eq. (1.13), the only momenta on which the hadronic tensor can depend are the momentum of the initial hadron, p_H, and the difference in momentum between the outgoing and ingoing electron, q; that is, $H_{\mu\nu} \equiv H_{\mu\nu}(p_H, q)$. The only Lorentz scalars which can be constructed from p_H and q are $p_H^2 = M_H^2$, which is a fixed constant, $q^2 = -Q^2$ and $p_H \cdot q = Q^2/2x$. It can then be shown (see e.g. Ref. [8]) that current conservation at the hadronic vertex implies that the most general Lorentz-covariant structure for the hadronic tensor is given by:

$$H_{\mu\nu}(p_H, q) \equiv \left(-\eta_{\mu\nu} + \frac{q_\mu q_\nu}{q^2}\right) F_1(x, Q^2) + \left(p_{H\mu} - \frac{p_H \cdot q}{q^2} q_\mu\right)\left(p_{H\nu} - \frac{p_H \cdot q}{q^2} q_\nu\right) \frac{F_2(x, Q^2)}{p_H \cdot q}, \tag{1.18}$$

for some scalar functions F_1, F_2, called the *neutral current DIS structure functions*.

Contracting with the expression for the leptonic tensor, Eq. (1.12),[4] we arrive at the following expression for the differential cross-section:

$$\frac{d^3\sigma}{dxdyd\phi} = \frac{2y\alpha^2}{Q^4}\left(2(p_\ell \cdot p_{\ell'})F_1 + \frac{1}{p_H \cdot q}[2(p_H \cdot p_\ell)(p_H \cdot p_{\ell'}) - M_H^2(p_\ell \cdot p_{\ell'})]F_2\right), \tag{1.19}$$

where $\alpha = e^2/4\pi$ is the fine structure constant of QED. To finish, we note that Eq. (1.15) allows us to write:

$$p_\ell \cdot p_{\ell'} = \frac{Q^2}{2}, \qquad p_H \cdot p_\ell = \frac{Q^2}{2xy}, \qquad p_H \cdot p_{\ell'} = \frac{Q^2}{2x}\left(\frac{1}{y} - 1\right). \tag{1.20}$$

[4]When performing this contraction, it is convenient to note that $q_\mu L^{\mu\nu} = 0$, by four-momentum conservation.

Overall then, we can re-express the differential cross-section as:

$$\frac{d^3\sigma}{dxdyd\phi} = \frac{2\alpha^2}{xyQ^2}\left(xy^2F_1 + \left[1 - y - M_H^2\frac{x^2y^2}{Q^2}\right]F_2\right). \tag{1.21}$$

We see that the right hand side possesses no dependence on the angular variable ϕ, so may be integrated directly to yield:

$$\frac{d^2\sigma}{dxdy} = \frac{4\pi\alpha^2}{xyQ^2}\left(xy^2F_1 + \left[1 - y - M_H^2\frac{x^2y^2}{Q^2}\right]F_2\right). \tag{1.22}$$

This is the final form for the (photon-mediated) DIS differential cross-section, in agreement with Eq. (19.8) of Ref. [3]. It is written in terms of the structure functions, F_1, F_2, which are the experimentally accessible observables; modelling the hadronic tensor therefore provides predictions for the observable quantities in DIS.

1.1.2 The parton model

So far, the hadronic tensor remains incalculable in perturbation theory, since it depends on the non-perturbative proton state $|H\rangle$. In order to model it, Feynman proposed a phenomenological model called the *parton model*, which we describe as follows.

It was originally argued in [7] that at ultra-relativistic energies (namely in the *deep inelastic limit* $Q^2 \to \infty$, where the energy transferred from the electron to the proton through the virtual photon approaches infinity), relativistic time-dilation results in the interactions in the proton happening over a characteristic scale $O(1/Q)$. In particular, this implies that at the moment of collision, the impacting electron will interact with only a *single* constituent of the proton. This suggests adopting the following phenomenological model, called the *parton model*, for the hadronic tensor:

$$H_{\mu\nu} = \frac{1}{4\pi}\sum_q \int_0^1 \frac{d\xi}{\xi}\sum_X \left(\prod_{i=1}^{n_X}\frac{d^{d-1}\mathbf{p}_{(i,X)}}{(2\pi)^{d-1}2E_{(i,X)}}\right)\cdot(2\pi)^d\delta^d\left(p_\ell + \xi p_H - p_{\ell'} - \sum_{i=1}^{n_X} p_{(i,X)}\right)$$
$$\cdot\,\overline{\langle q(\xi p_H)|J_\nu^\dagger|X\rangle\,\langle X|J_\mu|q(\xi p_H)\rangle} f_q(\xi), \tag{1.23}$$

Here, we have replaced the hadronic state $|H\rangle$ with a state $|q(\xi p_H)\rangle$; this denotes a state comprising a free constituent q of the proton,[5] carrying a fraction ξ of the momentum of the proton H. We sum over all possible constituents q, and we weight the contributions by some unknown, non-perturbative probability distributions $f_q(\xi)$; these distributions

[5]Note we also denote the virtuality of the mediating photon by q; this should cause no confusion as they enter in different ways.

represent the probability of the proton ejecting a constituent q to interact with the electron, carrying a momentum fraction ξ. Naturally, we also integrate over all possible momentum fractions. The bar over the amplitude product denotes the spin/colour-sum/average, as usual.

With this assumption on the form of the hadronic tensor, we may calculate the hadronic tensor at lowest order in QCD perturbation theory. In this case, we may take $|X\rangle$ to be a single quark state $|X\rangle = |q(p_X)\rangle$;[6] then, the hadronic tensor reduces to:

$$H_{\mu\nu}^{\mathrm{LO}} = \frac{1}{2}\sum_q \int_0^1 \frac{d\xi}{\xi} \int \frac{d^{d-1}\mathbf{p}_X}{2E_X}\delta^d\left(p_\ell + \xi p_H - p_{\ell'} - p_X\right) \overline{\langle q(\xi p_H)|J_\nu^\dagger|q(p_X)\rangle \langle q(p_X)|J_\mu|q(\xi p_H)\rangle} f_q(\xi). \tag{1.24}$$

Note that in this case, the colour sum/average of the amplitude product is trivial, since the quark does not change colour during the interaction. The phase space integral can be manipulated via:

$$\int \frac{d^{d-1}\mathbf{p}_X}{2E_X}\delta^d\left(p_\ell + \xi p_H - p_{\ell'} - p_X\right) = \int d^d p_X \delta(p_X^2)\delta^d\left(p_\ell + \xi p_H - p_{\ell'} - p_X\right), \tag{1.25}$$

which results in:

$$H_{\mu\nu}^{\mathrm{LO}} = \frac{1}{2}\sum_q \int_0^1 \frac{d\xi}{\xi}\ \delta((q+\xi p_H)^2)\cdot\ \overline{\langle q(\xi p_H)|J_\nu^\dagger|q(q+\xi p_H)\rangle \langle q(q+\xi p_H)|J_\mu|q(\xi p_H)\rangle} f_q(\xi). \tag{1.26}$$

In the deep inelastic limit $Q^2 \to \infty$, corresponding to very high energy transfer from the electron to the proton, we have that $(q+\xi p_H)^2 = q^2 + 2\xi q\cdot p_H + p_H^2 \approx q^2 + 2\xi q\cdot p_H$, and hence the delta function condition can be rewritten as:

$$\delta(q^2 + 2\xi q\cdot p_H) = \frac{\delta(x-\xi)}{2q\cdot p_H}, \tag{1.27}$$

where x is the Bjorken-x we introduced earlier in the text; this reveals that, at leading order in QCD perturbation theory, the interpretation of the Bjorken-x is the momentum fraction carried by the quark state ejected by the hadron which participates in the interaction with the photon. It follows that the hadronic tensor may be expressed in the parton model as:

$$H_{\mu\nu}^{\mathrm{LO}} = \frac{1}{4xq\cdot p_H}\sum_q \overline{\langle q(xp_H)|J_\nu^\dagger|q(q+xp_H)\rangle \langle q(q+xp_H)|J_\mu|q(xp_H)\rangle} f_q(x). \tag{1.28}$$

[6]Recall that X was initially defined to be a hadronic state; however, since we are summing over all hadronic states by completeness we may choose to use a basis of quarks and gluons rather than a basis of hadrons, similar to the discussion given around Eqs. (1.5) and (1.6).

We can now straightforwardly compute the matrix elements to yield:

$$\begin{aligned}
&\overline{\langle q(xp_H)|J_\nu^\dagger|q(q+xp_H)\rangle \langle q(q+xp_H)|J_\mu|q(xp_H)\rangle} \\
&\quad = \frac{1}{2}e_q^2 \mathrm{Tr}(x\not{p}_H\gamma_\nu(\not{q}+x\not{p}_H)\gamma_\mu) \\
&\quad = 2xe_q^2\left(p_{H\nu}(q+xp_H)_\mu + p_{H\mu}(q+xp_H)_\nu - \eta_{\mu\nu}p_H\cdot(q+xp_H)\right),
\end{aligned} \tag{1.29}$$

where e_q is the charge on the quark q in units of the electric charge e. Note the factor of $1/2$ coming from averaging the emitted quark spin states.

In the ultra-relativistic limit, it becomes appropriate to neglect the mass of the target proton, $M_H^2 \approx 0$; in this case, we can project the structure functions out of the hadronic tensor via the following contractions:

$$F_1(x,Q^2) = \left(-\frac{1}{2}\eta^{\mu\nu} + \frac{2x^2}{Q^2}p_H^\mu p_H^\nu\right)H_{\mu\nu}(p_H,q), \tag{1.30}$$

$$F_2(x,Q^2) = \left(-x\eta^{\mu\nu} + \frac{12x^3}{Q^2}p_H^\mu p_H^\nu\right)H_{\mu\nu}(p_H,q). \tag{1.31}$$

Applying the projectors, Eq. (1.30) and Eq. (1.31), we obtain the following formulae for the structure functions (to leading order in QCD):

$$F_1^{\mathrm{LO}}(x,Q^2) = \frac{1}{2}\sum_q e_q^2 f_q(x), \tag{1.32}$$

$$F_2^{\mathrm{LO}}(x,Q^2) = x\sum_q e_q^2 f_q(x). \tag{1.33}$$

Note the following important features of our final structure function formulae in the parton model:

(1) We have the relation $F_2^{\mathrm{LO}} = 2xF_1^{\mathrm{LO}}$. This identity is called the *Callan-Gross relation*, and provides evidence that the primary constituents of the proton have spin-$\frac{1}{2}$; if we assume different spins for the constituents of the proton, and attempt to perform the calculations presented in this section, we obtain different relations between $F_1^{\mathrm{LO}}, F_2^{\mathrm{LO}}$ which are not observed experimentally (indeed one obtains $F_1^{\mathrm{LO}} \equiv 0$ for scalar quarks, for example - see the discussion below Eq. (4.18) in [9]). The Callan-Gross relation is not *perfect*; deviations from this law are due to QCD corrections to the parton model, which we shall now discuss.

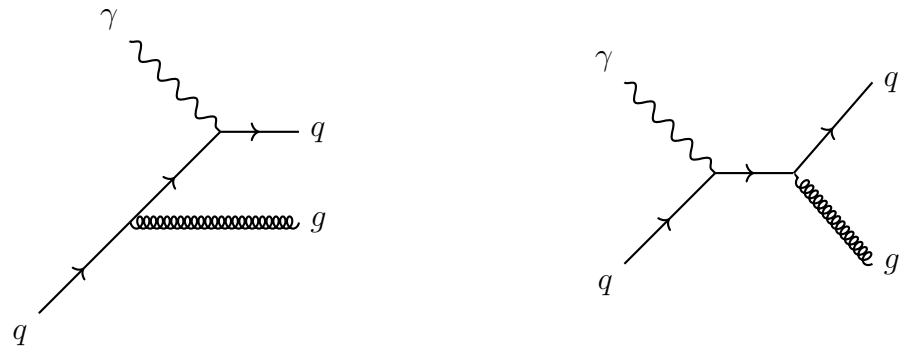


Figure 1.3: **Left.** Real gluon emission from the initial quark. **Right.** Real gluon emission from the final quark.

(2) The structure functions are independent of Q^2; this is called *Bjorken scaling*, and is experimentally observed when $Q^2 \to \infty$ (see for example Fig. 6 of Ref. [10]). Similarly, Bjorken scaling is broken by QCD corrections to the parton model.

1.1.3 The QCD-improved parton model

We can extend the parton model by including next-to-leading order QCD corrections. The diagrams which contribute to the next-to-leading order QCD corrections can be classified into two categories:

(i) **Real and virtual gluon emission.** In this case, a gluon is emitted from either the initial quark leg, or the final quark leg (shown on the left and right of Fig. 1.3 respectively). We take the final state to be $|X\rangle = |q(p_q)g(p_g)\rangle$, comprising a quark of four-momentum p_q and a gluon of four-momentum p_g.

Alternatively, a virtual gluon can be emitted from the initial or final quark leg and *reabsorbed*, rather than radiated. We will not give the details of the calculation of these diagrams for brevity, merely stating the results (we shall see that an important cancellation occurs between the real and virtual diagram contributions).

(ii) **Gluon-boson fusion.** In this case, a gluon is ejected from the proton, splitting into two quarks; one of the quarks is radiated, whilst the other participates in an interaction with the photon (see Fig. 1.4). Again, in the interest of being brief, we will not give the details of this calculation, simply stating the complete results at the end of the discussion.

We begin by computing the contribution to the hadronic tensor from (i), specifically focussing on real gluon emission. Throughout we work in d dimensions, so that we may employ dimensional regularisation when subsequently regulating the theory (we shall find that there are divergences in the calculation). The amplitudes for the diagrams in Fig. 1.3

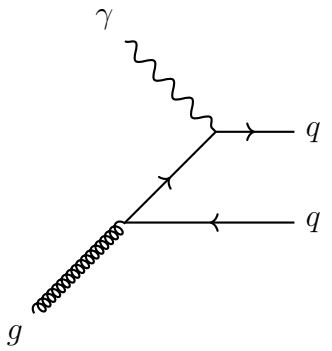


Figure 1.4: Ejection of a gluon from the proton, splitting into a quark which participates in the hard reaction with the photon, and a quark which is radiated. There is also a similar diagram involving antiquarks, where the arrows on the fermion lines are reversed.

are, respectively:

$$\mathcal{M}^i_\mu = -ig_S\mu^\epsilon e_q \bar{u}(p_q)\gamma_\mu \frac{1}{\xi \not{p}_H - \not{p}_g} \not{\epsilon}(p_g) t^A u(\xi p_H) \tag{1.34}$$

$$\mathcal{M}^f_\mu = -ig_S\mu^\epsilon e_q \bar{u}(p_q)\not{\epsilon}(p_g) \frac{1}{\not{p}_q + \not{p}_g} \gamma_\mu t^A u(\xi p_H), \tag{1.35}$$

where $g_S\mu^\epsilon$ is the coupling constant of QCD in d dimensions; this is the d-dimensional version of the coupling g_S appearing in Eq. (1.2), but now includes an additional factor of μ^ϵ where μ is an arbitrary mass scale and $\epsilon = \frac{4-d}{2}$. The polarisation of the gluon is given by $\epsilon_\mu(p_g)$. The factor of t^A denotes the relevant colour matrix from the Lie algebra $\mathfrak{su}_\mathbb{C}(3)$; technically this implies that $\mathcal{M}^i_\mu$ should also carry an $SU(3)$-index A corresponding to the colour state of the emitted gluon, but we will suppress this given that we will shortly sum/average over colours. The labels i, f on the respective amplitudes denote a gluon emitted from the $i$nitial quark and the $f$inal quark.

In this case, the contribution to the hadronic tensor is given by:

$$H^{\text{NLO, real}}_{\mu\nu} = \frac{1}{4\pi}\sum_q \int_0^1 \frac{d\xi}{\xi} f_q(\xi) \int \frac{d^{d-1}\mathbf{p}_q}{(2\pi)^{d-1}2E_q} \frac{d^{d-1}\mathbf{p}_g}{(2\pi)^{d-1}2E_g}$$
$$\cdot (2\pi)^d \delta^d\left(q + \xi p_H - p_q - p_g\right) \cdot \sum_{r,s=i,f} \overline{\mathcal{M}^r_\mu \mathcal{M}^{s\dagger}_\nu}, \tag{1.36}$$

where, as usual, the bar on the amplitude product denotes the spin/colour-sum/average. This expression can be evaluated in two steps: first, we will manipulate the phase space into a more amenable form, and then second, we shall compute the spin/colour-sum/averages of the amplitude products.

We begin by noting that:

$$\begin{aligned}
&\int \frac{d^{d-1}\mathbf{p}_q}{(2\pi)^{d-1}2E_q}\frac{d^{d-1}\mathbf{p}_g}{(2\pi)^{d-1}2E_g}\cdot(2\pi)^d\delta^d\left(q+\xi p_H-p_q-p_g\right)\\
&\quad=\frac{1}{2(2\pi)^{d-2}}\int\frac{d^{d-1}\mathbf{p}_g}{E_g}d^dp_q\;\delta(p_q^2)\delta^d\left(q+\xi p_H-p_q-p_g\right)\\
&\quad=\frac{1}{2(2\pi)^{d-2}}\int\frac{d^{d-1}\mathbf{p}_g}{E_g}\delta((q+\xi p_H-p_g)^2)
\end{aligned}\tag{1.37}$$

To absorb the remaining delta function, we make a sequence of variable changes. Consider first changing variables from $\mathbf{p}_g$ to $(|\mathbf{p}_g|^2,\cos(\theta))$, where θ is the angle defined by:

$$\cos(\theta)=\frac{\mathbf{p}_g\cdot\xi\mathbf{p}_H}{|\mathbf{p}_g||\xi\mathbf{p}_H|}=\frac{\mathbf{p}_g\cdot\mathbf{p}_H}{|\mathbf{p}_g||\mathbf{p}_H|}.\tag{1.38}$$

The measure transforms as:[7]

$$\begin{aligned}
\int\frac{d^{d-1}\mathbf{p}_g}{E_g}&=\frac{1}{2}\int|\mathbf{p}_g|^{d-4}\sin^{d-4}(\theta)d(|\mathbf{p}_g|^2)d(\cos(\theta))d\Omega_{d-3}\\
&=\frac{(d-2)\pi^{\frac{1}{2}(d-2)}}{2\Gamma(d/2)}\int|\mathbf{p}_g|^{d-4}\sin^{d-4}(\theta)d(|\mathbf{p}_g|^2)d(\cos(\theta))\\
&=\frac{\pi^{1-\epsilon}}{\Gamma(1-\epsilon)}\int|\mathbf{p}_g|^{d-4}\sin^{d-4}(\theta)d(|\mathbf{p}_g|^2)d(\cos(\theta)),
\end{aligned}\tag{1.39}$$

where $d\Omega_{d-3}$ contains all angular integrations in $d-3$ dimensions, except for the integration over θ; in the second line we perform the integration over $d\Omega_{d-3}$ using a standard hyperspherical integral, and in the final line we recall that $d=4-2\epsilon$.

We now make a second change of variables. Consider introducing the variables u,v defined by the conditions:

$$p_g\cdot p_q=\frac{Q^2}{2}\left(\frac{1-u}{u}\right),\qquad p_g\cdot\xi p_H=\frac{Q^2v}{2u}.\tag{1.40}$$

The interpretation of these variables will become clear shortly. In order to relate the variables u,v to the variables $|\mathbf{p}_g|^2$, $\cos(\theta)$, we work in the centre of momentum frame where $\mathbf{q}+\xi\mathbf{p}_H=\mathbf{0}=\mathbf{p}_g+\mathbf{p}_q$. The four-vectors $q,\xi p_H,p_g,p_q$ can then be expressed as:

$$q=\begin{pmatrix}\sqrt{\xi^2|\mathbf{p}_H|^2-Q^2}\\-\xi\mathbf{p}_H\end{pmatrix},\quad \xi p_H=\begin{pmatrix}\xi|\mathbf{p}_H|\\\xi\mathbf{p}_H\end{pmatrix},\quad p_g=\begin{pmatrix}|\mathbf{p}_g|\\\mathbf{p}_g\end{pmatrix},\quad p_q=\begin{pmatrix}|\mathbf{p}_g|\\-\mathbf{p}_g\end{pmatrix},\tag{1.41}$$

[7]Recall the gluon is real, so on-shell, and hence $E_g=|\mathbf{p}_g|$.

where the first component of the virtual photon's four-momentum can be computed by enforcing the condition $q^2 = -Q^2$. Using the definition of u from $p_g \cdot p_q$, and working in the above frame, we see that:

$$|\mathbf{p}_g|^2 = \frac{Q^2}{4}\left(\frac{1-u}{u}\right). \tag{1.42}$$

In particular, we see that u describes the *energy* of the radiated gluon; in particular, the limit $u \to 1$ corresponds to the limit in which the energy of the gluon goes to zero, the so-called *soft* limit.

To obtain a relation between v and $|\mathbf{p}_g|, \cos(\theta)$, we note that by conservation of energy, we can deduce that:

$$\sqrt{\xi^2|\mathbf{p}_H|^2 - Q^2} + \xi|\mathbf{p}_H| = 2|\mathbf{p}_g| \qquad \Rightarrow \qquad \xi|\mathbf{p}_H| = \frac{4|\mathbf{p}_g|^2 + Q^2}{4|\mathbf{p}_g|}. \tag{1.43}$$

Then the definition of v in terms of $p_g \cdot \xi p_H$, we have:

$$\begin{aligned}\frac{Q^2 v}{2u} &= |\mathbf{p}_g|\xi|\mathbf{p}_H| - \mathbf{p}_g \cdot \xi\mathbf{p}_H \\ &= |\mathbf{p}_g|\xi|\mathbf{p}_H|(1-\cos(\theta)) \\ &= \frac{Q^2}{4u}(1-\cos(\theta)),\end{aligned} \tag{1.44}$$

which yields:

$$v = \frac{1-\cos(\theta)}{2}. \tag{1.45}$$

Hence v is a measure of the *collinearity* of the outgoing quark and the radiated gluon; in particular, as $v \to 0$, we have that the outgoing quark and the radiated gluon are collinear.

Computing the Jacobian factor, we can now make the change of variables to u, v. We note that:

$$d(|\mathbf{k}_g|^2)d\cos(\theta) = \frac{\partial(|\mathbf{k}_g|^2, \cos(\theta))}{\partial(u,v)}dudv = \frac{Q^2}{2u^2}dudv, \tag{1.46}$$

and hence it follows that the phase space can be written in terms of u, v as:

$$\frac{1}{2(2\pi)^{2-2\epsilon}}\frac{\pi^{1-\epsilon}}{2\Gamma(1-\epsilon)}Q^{2(1-\epsilon)}\int dudv\ \frac{(1-u)^{-\epsilon}}{u^{2-\epsilon}}v^{-\epsilon}(1-v)^{-\epsilon}\delta((q+\xi p_H - p_g)^2). \tag{1.47}$$

To finish, we note that the argument of the delta function can be expanded to:

$$(q+\xi p_H - p_g)^2 = q^2 + 2q\cdot\xi p_H - 2q\cdot p_g - 2\xi p_H \cdot p_g. \tag{1.48}$$

Using Eq. (1.40), we can obtain the missing inner products of four-vectors which we require

in evaluating this expression. We find that:

$$(q+\xi p_H - p_g)^2 = \delta\left(Q^2\left(\frac{1}{u}-\frac{\xi}{x}\right)\right) = \frac{x\delta(\xi - x/u)}{Q^2}, \tag{1.49}$$

Using the delta function to absorb the ξ integral in the hadronic tensor, we see that we have reduced the hadronic tensor to:

$$H_{\mu\nu}^{\text{NLO, real}} = \frac{1}{4\pi}\sum_q \frac{1}{2(2\pi)^{2-2\epsilon}}\frac{\pi^{1-\epsilon}}{2\Gamma(1-\epsilon)}Q^{-2\epsilon}\int du dv\ f_q\left(\frac{x}{u}\right)\frac{(1-u)^{-\epsilon}}{u^{1-\epsilon}}v^{-\epsilon}(1-v)^{-\epsilon}\sum_{r,s=i,f}\overline{\mathcal{M}_\mu^r\mathcal{M}_\nu^{s\dagger}} \tag{1.50}$$

$$= \frac{1}{32\pi^2\Gamma(1-\epsilon)}\left(\frac{4\pi}{Q^2}\right)^\epsilon \sum_q \int_x^1 \frac{du}{u} f_q\left(\frac{x}{u}\right)\frac{u^\epsilon}{(1-u)^\epsilon}\int_0^1 dv\ v^{-\epsilon}(1-v)^{-\epsilon}\sum_{r,s=i,f}\overline{\mathcal{M}_\mu^r\mathcal{M}_\nu^{s\dagger}}. \tag{1.51}$$

The integration ranges can be obtained by considering the allowed regions of u, v arising from their definitions in Eq. (1.40).

It remains to compute the spin/colour-sum/averages of the quantities quadratic in the amplitudes, $\overline{\mathcal{M}_\mu^r\mathcal{M}_\nu^{s\dagger}}$. Here, we shall only compute $\overline{\mathcal{M}_\mu^i\mathcal{M}_\nu^{i\dagger}}$ to demonstrate the details of such a calculation; the others can be obtained by similar means. We begin by observing that:

$$\overline{\mathcal{M}_\mu^i\mathcal{M}_\nu^{i\dagger}} = \frac{e_q^2 g_S^2 \mu^{2\epsilon} C_F}{2(\xi p_H - p_g)^4} g_\perp^{\alpha\beta}(k_g)\text{Tr}\left(\not{p}_q\gamma_\mu(\xi\not{p}_H - \not{p}_g)\gamma_\alpha\xi\not{p}_H\gamma_\beta(\xi\not{p}_H - \not{p}_g)\gamma_\nu\right). \tag{1.52}$$

The trace arises from the usual method of taking spin-sum/averages of spinors. The factor of $C_F = 4/3$ arises from the colour-sum/averages; in particular, it is the quadratic Casimir of $\mathfrak{su}_\mathbb{C}(3)$, obeying:

$$\sum_A (t^A)^2 = C_F I. \tag{1.53}$$

Finally, the factor of $g_\perp^{\alpha\beta}(k_g)$ arises from the spin-sum/average of the radiated gluon; it is given by:

$$g_\perp^{\alpha\beta}(k_g) := \sum_{\text{polarisations}} \epsilon^\alpha(k_g)\epsilon^{\beta *}(k_g). \tag{1.54}$$

The sum depends on which polarisations are allowed in the final state. To simplify things as much as possible, we shall assume that *all* polarisations are allowed, including unphysical ones. This requires us to also consider the introduction of ghost fields; however, fortunately at this order in QCD there are no diagrams which include the ghosts. Therefore, we can

happily conclude that:

$$g_\perp^{\alpha\beta}(k_g) = \sum_{\text{polarisations}} \epsilon^\alpha(k_g)\epsilon^{\beta *}(k_g) = \eta^{\alpha\beta}, \tag{1.55}$$

the standard result when all polarisations are considered. This allows us to reduce Eq. (1.52) to the simplified form:

$$\begin{aligned}\overline{\mathcal{M}_\mu^i \mathcal{M}_\nu^{i\dagger}} &= \frac{e_q^2 g_S^2 \mu^{2\epsilon} C_F}{8(\xi p_H \cdot k_g)^2} \mathrm{Tr}\left(\not{p}_q \gamma_\mu (\xi \not{p}_H - \not{p}_g)\gamma_\alpha \xi \not{p}_H \gamma^\alpha (\xi \not{p}_H - \not{p}_g)\gamma_\nu\right) \\ &= \frac{2(1-\epsilon) e_q^2 g_S^2 \mu^{2\epsilon} C_F}{\xi p_H \cdot p_g} \left(\eta_{\mu\nu} p_g \cdot p_q - (p_g)_\mu (p_q)_\nu - (p_g)_\nu (p_q)_\mu\right), \end{aligned} \tag{1.56}$$

where the trace can be evaluated by standard Dirac algebra methods.[8] The factor of $1-\epsilon$ arises from a contraction of the form $\eta^{\alpha\beta}\eta_{\alpha\beta} = d = 4 - 2\epsilon$, recalling that we are working in $d = 4 - 2\epsilon$ dimensions.

Evaluating the other spin/colour-sum/averages of the quantities quadratic in the amplitudes, it can be shown that we have the following projections:

$$-\eta^{\mu\nu} \sum_{r,s=i,f} \overline{\mathcal{M}_\mu^r \mathcal{M}_\nu^{s\dagger}} = 4 e_q^2 g_S^2 \mu^{2\epsilon} C_F (1-\epsilon) \left[(1-\epsilon)\left(\frac{p_g \cdot p_q}{p_g \cdot \xi p_H} + \frac{p_g \cdot \xi p_H}{p_g \cdot p_q}\right) + \frac{Q^2(\xi p_H \cdot p_q)}{(p_g \cdot \xi p_H)(p_g \cdot p_q)} + 2\epsilon\right], \tag{1.57}$$

$$p_H^\mu p_H^\nu \sum_{r,s=i,f} \overline{\mathcal{M}_\mu^r \mathcal{M}_\nu^{s\dagger}} = \frac{4 e_q^2 g_S^2 \mu^{2\epsilon} C_F (1-\epsilon)}{\xi^2} (p_q \cdot \xi p_H). \tag{1.58}$$

These projections can be rewritten conveniently in terms of the 'softness' and 'collinearity' variables u, v we introduced in Eq. (1.40):

$$-\eta^{\mu\nu} \sum_{r,s=i,f} \overline{\mathcal{M}_\mu^r \mathcal{M}_\nu^{s\dagger}} = 4 e_q^2 g_S^2 \mu^{2\epsilon} C_F (1-\epsilon) \left[(1-\epsilon)\left(\frac{1-u}{v} + \frac{v}{1-u}\right) + \frac{2u(1-v)}{v(1-u)} + 2\epsilon\right], \tag{1.59}$$

$$p_H^\mu p_H^\nu \sum_{r,s=i,f} \overline{\mathcal{M}_\mu^r \mathcal{M}_\nu^{s\dagger}} = \frac{2 e_q^2 g_S^2 \mu^{2\epsilon} C_F (1-\epsilon) Q^2 (1-v) u}{x^2}. \tag{1.60}$$

[8] Note that in the case where a Z-boson or a W-boson mediates deep inelastic scattering, rather than a photon, these traces additionally contain γ^5 matrices. The definition of γ^5 in $d = 4 - 2\epsilon$ dimensions is rather subtle, and a lot more care is required in these calculations. A detailed discussion can be found in [11].

Inserting these formulae into the appropriate projections of Eq. (1.51), we obtain:

$$-\eta^{\mu\nu} H_{\mu\nu}^{\text{NLO, real}} = \frac{\alpha_S C_F (1-\epsilon)}{2\pi\Gamma(1-\epsilon)} \left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon} \sum_q e_q^2 \int_x^1 \frac{du}{u} f_q\left(\frac{x}{u}\right) \frac{u^{\epsilon}}{(1-u)^{\epsilon}}$$
$$\cdot \int_0^1 dv\ v^{-\epsilon}(1-v)^{-\epsilon} \left[(1-\epsilon)\left(\frac{1-u}{v} + \frac{v}{1-u}\right) + \frac{2u(1-v)}{v(1-u)} + 2\epsilon\right], \tag{1.61}$$

$$p_H^{\mu} p_H^{\nu} H_{\mu\nu}^{\text{NLO, real}} = \frac{\alpha_S Q^2 C_F (1-\epsilon)}{4\pi x^2 \Gamma(1-\epsilon)} \left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon} \sum_q e_q^2 \int_x^1 \frac{du}{u} f_q\left(\frac{x}{u}\right) \frac{u^{1+\epsilon}}{(1-u)^{\epsilon}} \int_0^1 dv\ v^{-\epsilon}(1-v)^{1-\epsilon}, \tag{1.62}$$

where $\alpha_S = g_S^2/4\pi$ is the strong coupling. We wish to determine the behaviour of these expressions as $\epsilon \to 0$, i.e. as we restore $d = 4$ dimensions. We note that the projection $p_H^{\mu} p_H^{\nu} H_{\mu\nu}^{\text{NLO, real}}$ remains finite in this limit, giving the result:

$$p_H^{\mu} p_H^{\nu} H_{\mu\nu}^{\text{NLO, real}} = \frac{\alpha_S Q^2 C_F}{8\pi x^2} \sum_q e_q^2 \int_x^1 du f_q\left(\frac{x}{u}\right). \tag{1.63}$$

On the other hand, the projection $-\eta^{\mu\nu} H_{\mu\nu}^{\text{NLO, real}}$ is singular as $\epsilon \to 0$, and requires a more careful treatment. To begin with, we note that we can evaluate the collinear v integral using the identity:[9]

$$\int_0^1 dv\ v^{p-1}(1-v)^{q-1} = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{1.64}$$

Applying this result repeatedly, we can simplify Eq. (1.61) to:

$$-\eta^{\mu\nu} H_{\mu\nu}^{\text{NLO, real}} = \frac{\alpha_S C_F (1-\epsilon)}{2\pi\Gamma(1-\epsilon)} \left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon} \sum_q e_q^2 \int_x^1 \frac{du}{u} f_q\left(\frac{x}{u}\right) \frac{u^{\epsilon}}{(1-u)^{\epsilon}}$$
$$\cdot \left[(1-\epsilon)\left(\frac{(1-u)\Gamma(-\epsilon)\Gamma(1-\epsilon)}{\Gamma(1-2\epsilon)} + \frac{\Gamma(2-\epsilon)\Gamma(1-\epsilon)}{(1-u)\Gamma(3-2\epsilon)}\right) + \frac{2u}{1-u}\frac{\Gamma(-\epsilon)\Gamma(2-\epsilon)}{\Gamma(2-2\epsilon)} + 2\epsilon\frac{\Gamma(1-\epsilon)^2}{\Gamma(2-2\epsilon)}\right], \tag{1.65}$$

which by repeated application of the functional equation of the gamma function, $\Gamma(x+1) =$

[9] Technically this result holds only when $\text{Re}(p), \text{Re}(q) > 0$, but we can always pretend that ϵ is an appropriate range before using analytic continuation to justify whichever final result we get.

$x\Gamma(x)$, may be reduced to the more convenient form:

$$-\eta^{\mu\nu}H_{\mu\nu}^{\text{NLO, real}} = \frac{\alpha_S C_F(1-\epsilon)\Gamma(1-\epsilon)}{2\pi\Gamma(1-2\epsilon)}\left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon}\sum_q e_q^2\int_x^1\frac{du}{u}f_q\left(\frac{x}{u}\right)\frac{u^{\epsilon}}{(1-u)^{\epsilon+1}}$$
$$\cdot\left[\left(-\frac{1}{\epsilon}+1\right)(1-u)^2+\frac{1-\epsilon}{2(1-2\epsilon)}-\frac{2u(1-\epsilon)}{\epsilon(1-2\epsilon)}+\frac{2\epsilon}{1-2\epsilon}\right]. \quad (1.66)$$

We must now Laurent expand this formula in small ϵ. The key result we shall require is:

$$\frac{u^{\epsilon}}{(1-u)^{\epsilon+1}} = -\frac{1}{\epsilon}\delta(1-u)+\left(\frac{1}{1-u}\right)_+-\epsilon\left(\frac{\log(1-u)}{1-u}\right)_++\epsilon\frac{\log(u)}{1-u}+O(\epsilon^2), \quad (1.67)$$

which holds only in a distributional sense.[10] A proof of this identity is given in App. B. Here, the + symbol denotes the *plus distribution*, defined by:

$$\int_0^1 F(u)_+G(u)\,du = \int_0^1 F(u)(G(u)-G(1))\,du. \quad (1.68)$$

Combining the identity Eq. (1.67) with the standard Laurent expansions of the rational function expression in the large square brackets, as $\epsilon\to 0$ we see that the projection $-\eta^{\mu\nu}H_{\mu\nu}^{\text{NLO, real}}$ behaves as:

$$-\eta^{\mu\nu}H_{\mu\nu}^{\text{NLO, real}} = \frac{\alpha_S C_F(1-\epsilon)\Gamma(1-\epsilon)}{2\pi\Gamma(1-2\epsilon)}\left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon}\sum_q e_q^2\int_x^1\frac{du}{u}f_q\left(\frac{x}{u}\right)\left[\left(\frac{2}{\epsilon^2}+\frac{3}{2\epsilon}+\frac{7}{2}\right)\delta(1-u)\right.$$
$$\left.-\frac{1}{\epsilon}\frac{1+u^2}{(1-u)_+}+3-u-\frac{3}{2(1-u)_+}+(1+u^2)\left(\frac{\log(1-u)}{1-u}\right)_+-\frac{1+u^2}{1-u}\log(u)+O(\epsilon)\right]. \quad (1.69)$$

We see that there is a double pole as $\epsilon\to 0$; the origin of this can be traced back to the soft singularity as $u\to 1$ (this is made manifest in Eq. (1.67)), combined with the collinear singularity as $v\to 0$ (this is made manifest in Eq. (1.66), since after we have performed the v integration, we obtain a $1/\epsilon$ term).

We must sum the real gluon emissions with the virtual gluon emissions, which enter at the same order in QCD, and multiply the same quark distributions $f_q(\xi)$ in the parton model for the hadronic tensor. The calculation of these diagrams is equally long and arduous,[11]

[10] That is, when integrated against a smooth function from $u=0$ to $u=1$.

[11] Though it is possible to use some physical arguments to reduce the workload, see e.g. the discussion around Eq. (4.70) of Ref. [9].

so we shall merely state the complete results here. We find that in the limit as $\epsilon \to 0$, we obtain the contribution:

$$-\eta^{\mu\nu} H_{\mu\nu}^{\text{NLO, virtual}} = -\frac{\alpha_S C_F (1-\epsilon)\Gamma(1-\epsilon)}{2\pi\Gamma(1-2\epsilon)} \left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon} \sum_q e_q^2 \int_x^1 \frac{du}{u} f_q\left(\frac{x}{u}\right) \delta(1-u) \cdot \left(\frac{2}{\epsilon^2} + \frac{3}{\epsilon} + 8 + \frac{\pi^2}{3} + O(\epsilon)\right) \tag{1.70}$$

with $p_H^\mu p_H^\nu H_{\mu\nu}^{\text{NLO, virtual}}$ vanishing as $\epsilon \to 0$. We note that the double pole, the $O(\epsilon^{-2})$ term, is cancelled exactly between the real and virtual corrections.

Putting everything together then, we obtain the following complete formulae for the projections of the quark-initiated part of the hadronic tensor at next-to-leading order in QCD perturbation theory:

$$\begin{aligned} -\eta^{\mu\nu} H_{\mu\nu}^{\text{quark, LO+NLO}} &= \sum_q e_q^2 f_q(x) + \frac{\alpha_S C_F (1-\epsilon)\Gamma(1-\epsilon)}{2\pi\Gamma(1-2\epsilon)} \left(\frac{4\pi\mu^2}{Q^2}\right)^{\epsilon} \sum_q e_q^2 \int_x^1 \frac{du}{u} \\ &\quad \cdot f_q\left(\frac{x}{u}\right) \left[\left(-\frac{3}{2\epsilon} - \frac{\pi^2}{3} - \frac{9}{2}\right) \delta(1-u) - \frac{1}{\epsilon} \frac{1+u^2}{(1-u)_+} + 3 - u \right. \\ &\quad \left. - \frac{3}{2(1-u)_+} + (1+u^2) \left(\frac{\log(1-u)}{1-u}\right)_+ - \frac{1+u^2}{1-u} \log(u) \right], \end{aligned} \tag{1.71}$$

$$p_H^\mu p_H^\nu H_{\mu\nu}^{\text{quark, LO+NLO}} = \frac{\alpha_S C_F Q^2}{8\pi x^2} \sum_q e_q^2 \int_x^1 du \; f_q\left(\frac{x}{u}\right). \tag{1.72}$$

Rather disturbingly, we have discovered that in the limit $\epsilon \to 0$, the hadronic tensor in the parton model at next-to-leading order in QCD appears to be *singular*, corresponding to a collinear divergence from the real gluon emission that has not been cancelled with the divergence arising from the virtual gluon emission. It is at this point that the distributions $f_q(\xi)$ begin their starring role.

1.1.4 Definition of the $\overline{\text{MS}}$ PDFs

We initially introduced $f_q(\xi)$ as a probability distribution, representing the probability of a quark of momentum fraction ξ being ejected from the proton and participating in the

interaction with the photon. However, just as in the process of ultraviolet renormalisation, we can instead drop this interpretation beyond the leading order,[12] and instead view these distributions as 'bare theory parameters', subsequently renormalising them to absorb the leftover collinear divergences. In particular, this can be effected by allowing the 'bare distributions' $f_q(\xi)$ to have an ϵ dependence, $f_q(\xi) \equiv f_q(\xi, \epsilon)$. If the bare distribution $f_q(\xi, \epsilon)$ is divergent as $\epsilon \to 0$, then it is no longer necessary that the hadronic tensor projection in Eq. (1.71) is singular as $\epsilon \to 0$, since the divergences may now cancel one another.

Hence, with a view to redefining the 'bare distributions' $f_q(x, \epsilon)$ to absorb all divergences in Eq. (1.71), let us make the definition:

$$f_q^{\overline{\text{MS}}}(x, \mu_F^2) := f_q(x, \epsilon) - \frac{\alpha_S}{2\pi} \int_x^1 \frac{du}{u} f_q\left(\frac{x}{u}, \epsilon\right) P_{qq}(u) \left(\frac{1}{\epsilon} - \log\left(\frac{\mu_F^2}{\mu^2}\right) - \gamma + \log(4\pi)\right), \tag{1.73}$$

where γ is the Euler-Mascheroni constant,[13] and $P_{qq}(u)$ is called the *quark-quark splitting function*[14] defined by:

$$P_{qq}(u) := C_F \left[\frac{3}{2}\delta(1-u) + \frac{1+u^2}{(1-u)+}\right]. \tag{1.74}$$

The new object we have defined here is called the *modified minimal subtraction NLO quark parton distribution function* (which we shall henceforth refer to simply as an NLO *parton distribution function* (NLO PDF), or even just PDF when the order is implied; no other subtraction scheme will be considered in this thesis). It is a renormalised version of the 'bare PDF' $f_q(x, \epsilon)$. The term 'modified minimal subtraction' refers to this redefinition purely absorbing the collinear divergence, i.e. the pole term of order $O(1/\epsilon)$, plus the universal terms $-\gamma$, $\log(4\pi)$, which come from the expansion of the gamma function and the power law $(4\pi Q^2/\mu^2)^\epsilon$.

[12] Just as the 'bare mass' in a QFT Lagrangian no longer has the interpretation of mass beyond tree level.

[13] Which appears in the Taylor expansion $\Gamma(1+x) = 1 - \gamma x + O(x^2)$ about $x = 0$.

[14] Whilst the standard name is *function*, $P_{qq}(u)$ is technically a *distribution*, and only makes sense when integrated against a smooth function.

This redefinition allows us to rewrite the projection $-\eta^{\mu\nu} H_{\mu\nu}^{\text{quark, LO+NLO}}$ as:

$$-\eta^{\mu\nu} H_{\mu\nu}^{\text{quark, LO+NLO}} = \sum_q e_q^2 f_q^{\overline{\text{MS}}}(x,\mu_F^2) + \frac{\alpha_S C_F}{2\pi} \sum_q e_q^2 \int_x^1 \frac{du}{u}\, f_q^{\overline{\text{MS}}}\left(\frac{x}{u},\mu_F^2\right) \cdot \left[\frac{P_{qq}(u)}{C_F} \log\left(\frac{Q^2}{\mu_F^2}\right)\right.$$
$$\left. - \left(\frac{\pi^2}{3} + \frac{9}{2}\right)\delta(1-u) + 3 - u - \frac{3}{2(1-u)_+} + (1+u^2)\left(\frac{\log(1-u)}{1-u}\right)_+ - \frac{1+u^2}{1-u}\log(u)\right] + O(\alpha_S^2), \tag{1.75}$$

which is now finite in the limit as $\epsilon \to 0$. The parameter μ_F^2 in the parton distributions is referred to as the *factorisation scale*, and controls the finite part which is absorbed by the redefinition along with the singular part, which is of course arbitrary. However, it is customary to take $\mu_F^2 = Q^2$ to cancel any large logarithmic contribution coming from the first term in the NLO contribution of Eq. (1.75).

We also note that the parton distributions $f_q^{\overline{\text{MS}}}(x,\mu_F^2)$ are themselves now finite in the limit as $\epsilon \to 0$; since the projection $-\eta^{\mu\nu} H_{\mu\nu}^{\text{quark, LO+NLO}}$ is physical and hence finite, and all the functions that the PDF multiplies on the right hand side of Eq. (1.75) are now finite, it follows that $f_q^{\overline{\text{MS}}}(x,\mu_F^2)$ must be finite too. This implies a cancellation between the divergent behaviour of the bare PDF $f_q(x,\epsilon)$ as $\epsilon \to 0$ and the collinear pole $1/\epsilon$ as $\epsilon \to 0$ (at least at this order in perturbation theory in α_S).

1.1.5 Complete results for the NLO DIS structure functions

Combining with the gluon-initiated diagrams (ii), and allowing for a slightly more involved redefinition of f_q, f_g which now involves quark-gluon mixing, we obtain final expressions for the hadronic tensor at next-to-leading order in QCD perturbation theory. When the structure functions are finally projected out using Eq. (1.30) and Eq. (1.31), we obtain

the final results:

$$F_2^{\text{LO+NLO}} = x\sum_q e_q^2 \int_x^1 \frac{du}{u}\; f_q^{\overline{\text{MS}}}\left(\frac{x}{u},\mu_F^2\right)\left(\delta(1-u) + \frac{\alpha_S}{2\pi}P_{qq}(u)\log\left(\frac{Q^2}{\mu_F^2}\right)\right.$$
$$\left. + \frac{\alpha_S C_F}{2\pi}\left[\frac{1+u^2}{1-u}\left(\frac{\log(1-u)}{u} - \frac{3}{4}\right) + \frac{5u+9}{4}\right]_+\right)$$
$$+ x\sum_q e_q^2 \int_x^1 \frac{du}{u}\; f_g^{\overline{\text{MS}}}\left(\frac{x}{u},\mu_F^2\right)\frac{\alpha_S}{2\pi}\left(P_{qg}(u)\log\left(\frac{Q^2}{\mu_F^2}\right)\right.$$
$$\left. + T_R\left[(u^2+(1-u)^2)\log\left(\frac{1-u}{u}\right) - 1 + 8u(1-u)\right]\right), \quad (1.76)$$

$$F_1^{\text{LO+NLO}} = \frac{1}{2x}F_2^{\text{LO+NLO}} - \frac{\alpha_S}{2\pi}\sum_q e_q^2 \int_x^1 \frac{du}{u}\left(C_F f_q^{\overline{\text{MS}}}\left(\frac{x}{u}\right)u + 4T_R f_g^{\overline{\text{MS}}}u(1-u)\right). \quad (1.77)$$

where the splitting function $P_{qg}(u)$ is defined by:

$$P_{qg}(u) := T_R(u^2+(1-u)^2), \quad (1.78)$$

with $T_R = 1/2$ a Casimir of a further $\mathfrak{su}_{\mathbb{C}}(3)$ representation, and where for compactness we have applied the distributional identity:

$$-\left(\frac{\pi^2}{3}+\frac{9}{2}\right)\delta(1-u) + 3 + 2u - \frac{3}{2(1-u)_+} + (1+u^2)\left(\frac{\log(1-u)}{1-u}\right)_+ - \frac{1+u^2}{1-u}\log(u)$$
$$\equiv \left[\frac{1+u^2}{1-u}\left(\frac{\log(1-u)}{u} - \frac{3}{4}\right) + \frac{5u+9}{4}\right]_+, \quad (1.79)$$

which is proved in App. B. We can further simplify notation by introducing the *Mellin convolution*:

$$(f\otimes g)(x) := \int_x^1 \frac{du}{u} f(u) g\left(\frac{x}{u}\right) = \int_x^1 \frac{du}{u} f\left(\frac{x}{u}\right) g(u), \quad (1.80)$$

in terms of which the final next-to-leading order structure function formulae can be written as:

$$F_2^{\text{LO+NLO}} = \sum_q \hat{F}_2^{q,\text{LO+NLO}} \otimes f_q^{\overline{\text{MS}}} + \hat{F}_2^{g,\text{LO+NLO}} \otimes f_g^{\overline{\text{MS}}} \tag{1.81}$$

$$F_1^{\text{LO+NLO}} = \frac{1}{2x} F_2^{\text{LO+NLO}} - \frac{\alpha_S}{2\pi} \sum_q e_q^2 \left(C_F f_q^{\overline{\text{MS}}} \otimes u + 4T_R f_g^{\overline{\text{MS}}} \otimes u(1-u) \right), \tag{1.82}$$

where we have the 'partonic' structure function expressions:

$$\begin{aligned} \hat{F}_2^{q,\text{LO+NLO}} = x e_q^2 \Bigg(& \delta(1-u) + \frac{\alpha_S}{2\pi} P_{qq}(u) \log\left(\frac{Q^2}{\mu_F^2}\right) \\ & + \frac{\alpha_S C_F}{2\pi} \left[\frac{1+u^2}{1-u} \left(\frac{\log(1-u)}{u} - \frac{3}{4} \right) + \frac{5u+9}{4} \right]_+ \Bigg), \end{aligned} \tag{1.83}$$

$$\begin{aligned} \hat{F}_2^{g,\text{LO+NLO}} = \frac{x\alpha_S}{2\pi} \Bigg(& P_{qg}(u) \log\left(\frac{Q^2}{\mu_F^2}\right) \\ & + T_R \left[(u^2 + (1-u)^2) \log\left(\frac{1-u}{u}\right) - 1 + 8u(1-u) \right] \Bigg). \end{aligned} \tag{1.84}$$

In particular, Eq. (1.81) and Eq. (1.82) are of the 'factorisation theorem' form described in Eq. (1.7), where in this case we have a single PDF and no fragmentation functions. Whilst we have not rigorously *proved* factorisation (some notes on various ways in which this can be achieved are given at the end of this section), the work we have done in constructing Eqs. (1.81) and (1.82) lends credence to the idea that factorisation is true.

1.1.6 Universality of PDFs

Before concluding this section with a discussion of how factorisation theorems can be rigorously *proved*, we make an important remark regarding the *universality* of the PDFs constructed in this section.

Whilst the NLO $\overline{\text{MS}}$ PDFs we defined in this section were introduced in a study of DIS, their definition actually applies in a wide variety of processes where hadrons are present in the initial state. Indeed, we can see this from the definition Eq. (1.73), since the only part that relied on any amplitude calculation was the *splitting function* $P_{qq}(u)$. Furthermore, the part of the amplitude calculation that goes into the computation of the splitting function $P_{qq}(u)$ can be shown to be independent of the fact we are studying DIS; indeed, it can be shown (see e.g. the discussion in Altarelli and Parisi's famous paper, Ref. [12]) to be the probability that a quark radiates a collinear quark and a collinear gluon, with the quark carrying a fraction u of the initial quark's momentum. This applies

similarly to the splitting function $P_{qg}(u)$ we defined above; in general, we can construct splitting functions $P_{ij}(u)$ whose interpretation is the probability that a parton of type j emits a collinear parton of type i, with momentum fraction u of the original parton (possibly with other partons as appropriate).

Thus, since the NLO $\overline{\text{MS}}$ PDFs are constructed purely from universal, process-independent objects, they themselves are universal. For example, it can be shown that for the *Drell-Yan process*, $pp \to \ell^+\ell^- X$, involving two protons colliding to produce a detected lepton-antilepton pair and any other hadronic state X, we obtain the following factorisation theorem for the cross-section:

$$\sigma = \sum_{q_1,q_2} \hat{\sigma} \otimes f_{q_1}^{\overline{\text{MS}}} \otimes f_{q_2}^{\overline{\text{MS}}}, \tag{1.85}$$

where the sum over q_1, q_2 is over all proton constituents (including gluons), $\hat{\sigma}$ is the relevant perturbatively-calculable hard cross-section, and $f_{q_i}^{\overline{\text{MS}}}$ are the *same* PDFs we defined in the DIS construction above. This is the great attraction of factorisation theorems - we have packaged all of our ignorance of non-perturbative physics into universal objects which can be deployed in a very wide variety of processes, leaving only the perturbative, hard physics to calculate on a case by case basis.

1.1.7 Proofs of factorisation

Despite the intuitive picture that the parton model and its QCD improvement provide, we have not in any sense *proved* that our factorisation theorem is valid; in the above, we merely introduced a phenomenological model for the hadronic tensor, Eq. (1.23), essentially conjecturing its form in terms of unknown functions.

It is possible to rigorously derive the parton model from first principles in QCD, but the proofs are difficult and are beyond the scope of this thesis. In the case of DIS, there is a (somewhat) elementary proof relying on Wilson's *operator product expansion* (OPE), and explained in detail in Chapter 32.4.3 of Ref. [13]). However, the use of the OPE is limited, and more sophisticated techniques are required for more general processes than DIS.

One of the most successful methods of proving factorisation theorems is the method of Collins, Soper and Sterman (summarised nicely in Ref. [14]), which involves showing that the dominant Feynman diagrams for a given process can be organised into a 'hard' sub-diagram and a 'PDF' sub-diagram (these need not be the case naïvely - we could have double parton emission, or even more complicated diagrams, hence we must show that such diagrams are negligible[15]). The method is reasonably accessible in the case of a scalar

[15]In fact, they are shown to be suppressed by powers of a characteristic energy scale of the process; these contributions are called *higher twist* in the literature.

theory, where the hadron is supposed to be built of spin-0 constituents (see Chapter 5 of Ref. [14] for a full explanation), but in fully-fledged QCD becomes more difficult; the reason is that collinear divergences are also joined by *soft* divergences corresponding to the massless gluons having zero energy - showing that the soft divergences cancel is extremely subtle.

More recently, factorisation theorems have been proved in the *soft-collinear effective field theory* (SCET) approach (see Ref. [15], for example). This relies on constructing an effective theory of QCD, where the separation of the theory is based on collinearity and softness (rather than in a standard effective theory, which involves integrating out massive particles). We shall have a lot more to say about effective field theories (EFTs) in Chapters 3 and 4.

1.2 Properties of parton distribution functions

Having defined PDFs (at least at NLO in QCD), it will be useful to report some of their salient properties; in this section, we describe some key features of the PDFs which shall be used throughout this thesis. First, we construct the *evolution equations* for parton distributions in the $\overline{\text{MS}}$ scheme; these evolution equations give a perturbative description of the dependence of the the PDFs on the factorisation scale. Second, we state the *valence* and *momentum* sum rules for PDFs. Finally, we comment on two important features of the PDFs: their *positivity*, and their large-x and small-x *scaling* behaviour.

1.2.1 Evolution equations

Consider taking the logarithmic μ_F^2 derivative of Eq. (1.73) above. We find that:

$$\begin{aligned}\mu_F^2 \frac{\partial}{\partial \mu_F^2} f_q^{\overline{\text{MS}}}(x, \mu_F^2) &= \frac{\alpha_S}{2\pi} \int_x^1 \frac{du}{u} P_{qq}(u) f_q^{\overline{\text{MS}}}\left(\frac{x}{u}, \mu_F^2\right) + O(\alpha_S^2) \\ &= \frac{\alpha_S}{2\pi} (P_{qq} \otimes f_q^{\overline{\text{MS}}})(x, \mu_F^2) + O(\alpha_S^2). \end{aligned} \tag{1.86}$$

In particular, we see that the quark parton distribution functions obey an integro-differential equation in their second argument; this equation is called the *Dokshitzer-Gribov-Lipatov-Altarelli-Parisi* (DGLAP) equation [12, 16, 17], and is the analogue of a Callan-Symanzik equation from standard ultraviolet renormalisation. Of course, Eq. (1.86) ignores the gluon distribution; if we make the redefinition in Eq. (1.73) also including the gluon-initiated diagrams, we must allow for quark-gluon mixing, which leads to the more complete set of

DGLAP equations:

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} \begin{pmatrix} f_q^{\overline{\text{MS}}}(x,\mu_F^2) \\ f_g^{\overline{\text{MS}}}(x,\mu_F^2) \end{pmatrix} = \frac{\alpha_S}{2\pi} \int_x^1 \frac{du}{u} \begin{pmatrix} P_{qq}(u) & P_{qg}(u) \\ P_{gq}(u) & P_{gg}(u) \end{pmatrix} \begin{pmatrix} f_q^{\overline{\text{MS}}}(x/u,\mu_F^2) \\ f_g^{\overline{\text{MS}}}(x/u,\mu_F^2) \end{pmatrix} + O(\alpha_S^2), \quad (1.87)$$

$$= \frac{\alpha_S}{2\pi} \begin{pmatrix} P_{qq} & P_{qg} \\ P_{gq} & P_{gg} \end{pmatrix} \otimes \begin{pmatrix} f_q^{\overline{\text{MS}}} \\ f_g^{\overline{\text{MS}}} \end{pmatrix} (x,\mu_F^2) + O(\alpha_S^2), \quad (1.88)$$

where the relevant additional splitting functions are given by:

$$P_{gq}(u) := C_F \left(\frac{1+(1-u)^2}{u} \right), \quad (1.89)$$

$$P_{gg}(u) := 2C_A \left(\frac{u}{(1-u)_+} + \frac{1-u}{u} + u(1-u) \right) + \delta(1-u) \frac{(11C_A - 24T_R)}{6}, \quad (1.90)$$

with $C_A = 3$ the Casimir of a further representation of $\mathfrak{su}_{\mathbb{C}}(3)$. As described above, these splitting functions can be determined in a process-independent way; this will be particularly useful in Chapter 2.

It is also important to note that the equations we have presented above are accurate to next-to-leading order in QCD perturbation theory; we can extend these equations using higher order QCD splitting functions, and indeed QED splitting functions, which shall be relevant in Chapter 2. The QCD contributions to the splitting functions were fully computed up to $O(\alpha_S^3)$ in [18, 19, 20],[16] the mixed QED and QCD contribution was computed in [22], and the NLO QED contribution was computed in [23].

These evolution equations can be solved numerically. Software able to do this includes APFEL (**A** P**D**F **E**volution **L**ibrary) [24] and EKO (**E**volution **K**ernel **O**perators) [25]; the former is used throughout the thesis with occasional modification. In both cases, a rotation of PDF flavours is made to simplify the matrix of splitting functions, decoupling some of the evolution equations from one another; more details are presented in [24], with some further discussion in Chapter 2.

1.2.2 Sum rules

The initial interpretation of the 'bare' PDFs as probability distributions implies that they should obey certain '*sum rules*'. In particular, since hadrons are composed of a collection of a fixed number of *valence quarks*, together with particles generated by virtual exchange,

[16] They are also partially known at $O(\alpha_S^4)$ [21], but this contribution are not yet fully known and are not included in any public PDF evolution code.

we must have:

$$\int_0^1 dx \ \left(f_q(x) - f_{\bar{q}}(x)\right) = N_q, \tag{1.91}$$

for each of the quarks q, where $\bar{q}$ denotes the associated antiquark, and where N_q is the number of valence quarks of type q which form the hadron. In the case of a proton, we have $N_u = 2$ and $N_d = 1$, with $N_q = 0$ for all other flavours of quarks. This relation survives the redefinition of the PDFs, Eq. (1.73), yielding the *valence sum rules*:

$$\int_0^1 dx \ \left(f_q^{\overline{\text{MS}}}(x, \mu_F^2) - f_{\bar{q}}^{\overline{\text{MS}}}(x, \mu_F^2)\right) = N_q. \tag{1.92}$$

Similarly, the introduction of the PDFs as probability distributions in momentum fraction, x, means that we have the *momentum sum rule*:

$$\int_0^1 dx \ x \left(\sum_q f_q^{\overline{\text{MS}}}(x, \mu_F^2) + f_g^{\overline{\text{MS}}}(x, \mu_F^2)\right) = 1, \tag{1.93}$$

where the sum over q is over all quarks and antiquarks.

These rules are naturally extended to accommodate additional constituents of the proton; for example, if we study the proton at higher order in QED, the momentum sum rule must be adjusted to accommodate a contribution from the *photon* PDF, and the *lepton* PDFs. These considerations will be important in Chapter 2.

1.2.3 Positivity

When we introduced PDFs in Sect. 1.1 in the context of the parton model, they were motivated as probability densities describing the probability of ejecting different constituents of the proton carrying different momentum fractions; as such, the 'bare' distributions should in principle be *positive* quantities. However, we saw that the inclusion of NLO QCD effects requires a redefinition of these 'bare' distributions, given in Eq. (1.73), removing this initial interpretation; therefore, traditionally there has been no expectation for PDFs beyond the leading order to be positive.

However, a proof has recently become available that this is indeed the case [26]; the proof is beyond the scope of this thesis, but we shall assume its result - namely that for each PDF flavour, we have:

$$f^{\overline{\text{MS}}}(x, \mu_F^2) \geq 0 \tag{1.94}$$

for all $x \in [0, 1], \mu_F^2 \in [0, \infty)$.[17]

[17]It is worth noting, however, that the rigour of the proof is somewhat in question; in particular,

1.2.4 Large-x and small-x behaviour

Since a constituent of the proton cannot carry a momentum fraction $x > 1$, we must have:

$$\lim_{x\to 1} f^{\overline{\text{MS}}}(x, \mu_F^2) = 0 \tag{1.95}$$

for all flavours of PDF. On the other hand, the scaling behaviour of the PDFs in the limit as $x \to 0$ is also known, dictated by *Regge theory*; in brief, this theory tells us that, quite generally, scattering amplitude are proportional to certain power laws governed by information on the angular momentum of the process. It turns out (see Ref. [29]) that this implies the PDFs themselves must obey a power law scaling in the small-x limit:

$$f^{\overline{\text{MS}}}(x, \mu_F^2) \propto x^{\alpha}, \qquad \text{as } x \to 0, \tag{1.96}$$

for each flavour of PDF (though α will depend on the flavour). The upshot of these two scaling limits is that PDFs take the general form:

$$f^{\overline{\text{MS}}}(x, \mu_F^2) = x^{\alpha}(1-x)^{\beta}\tilde{f}(x, \mu_F^2), \tag{1.97}$$

with $\beta > 0$, for some unknown function $\tilde{f}$; this suggests the functional form that most PDF fitting collaborations build upon, as we shall discuss in the subsequent section, Sect. 1.3.

1.3 Fitting parton distribution functions

Above, we introduced parton distributions to parametrise the non-perturbative structure of hadrons. Their very definition implies that they cannot be obtained by perturbative methods, which essentially leaves open two options for their determination: (i) lattice methods (see e.g. [30] for some progress in this direction); (ii) fits to experimental data. In this text, we focus exclusively on the latter choice, which we shall describe in detail in this section.

We begin by describing possible functional forms which can be used to model the PDFs. The parameters in these functional forms are obtained by fits to a given dataset by the minimisation of a loss function, namely the χ^2-statistic (in the t_0 prescription, with various penalty terms), which we subsequently motivate and define. Next, we describe a method for minimisation of the loss function, namely *stochastic gradient descent*. Finally, we describe a standard method of error analysis in PDF fits, namely the *Monte Carlo*

Ref. [27] shows that in fact 'bare' PDFs need not be positive in $d = 4 - 2\epsilon$ dimensions, which is a major assumption of the proof in Ref. [26]. A response was given by the NNPDF collaboration in [28], and it is on this basis that the NNPDF fitting collaboration assume positivity of the PDFs; given that this thesis works closely with their methodology, we shall do the same.

replica method, which allows for the propagation of experimental uncertainties onto the PDFs.

A discussion of the actual datasets used in PDF fits is deliberately omitted and is deferred to later chapters; different datasets will be used to fit the PDFs for each of the scenarios considered in this text.

1.3.1 The choice of functional form

The space of possible parton distributions is an *infinite*-dimensional function space, comprising solutions of the DGLAP equations, Eq. (1.88), which obey the momentum and valence sum rules; as such, given only *finite* amounts of data, it is an ill-posed problem to determine the PDFs. Therefore, any PDF fitting attempt must initially restrict the infinite-dimensional PDF space to a *finite*-dimensional space instead, by assuming a functional form for the PDFs. Typically, a functional form is assumed at some initial factorisation scale $Q = Q_0$, below any characteristic energy scale for data entering the fit, and then DGLAP evolution is used to obtain the PDF at all scales. For the modern datasets used by most PDF fitting collaborations, the initial scale for fits is typically taken to be $Q_0 = 1.65$ GeV (as in, say, Ref. [31]).

Many functional forms are available; to give an early example, in Martin, Stirling and Roberts' analysis of PDFs in 1994 [10], the authors choose the following parametrisation for the valence quark distributions and the gluon distribution:[18]

$$x(f_u - f_{\bar{u}})(x, Q_0^2) = A_u x^{\eta_1}(1-x)^{\eta_2}(1 + \epsilon_u\sqrt{x} + \gamma_u x), \tag{1.98}$$

$$x(f_d - f_{\bar{d}})(x, Q_0^2) = A_d x^{\eta_3}(1-x)^{\eta_4}(1 + \epsilon_d\sqrt{x} + \gamma_d x), \tag{1.99}$$

$$xf_g(x, Q_0^2) = A_g x^{-\lambda}(1-x)^{\eta_g}(1 + \gamma_g x). \tag{1.100}$$

This form is motivated by the known large-x and small-x scaling behaviour of the PDFs, which we described above in Sect. 1.2.4, supplemented by a polynomial factor in $\sqrt{x}$ (the choice of a polynomial in $\sqrt{x}$ rather than in x is found to give a better fit). The authors also impose certain flavour assumptions, for example that the strange content of the proton is equal to the anti-strange content of the proton ($f_s = f_{\bar{s}}$), due to the datasets not being adequately able to disentangle the two.

Fitting groups have since extended this basic functional form and removed flavour assumptions as more data has become available, allowing the PDFs to be determined more precisely. Some modern fitting collaborations, for example the *The Coordinated Theoretical-Experimental Project on QCD* (CTEQ) group, even use an ensemble of functional forms to account for the bias expected by restricting to a particular individual functional form

[18] For brevity, we shall now drop the superscript $\overline{\text{MS}}$ denoting the modified minimal subtraction PDFs; we shall henceforth assume that all PDFs use this scheme.

(see Ref. [32], for example).

More recently, the *Neural Network Parton Distribution Function* (NNPDF) collaboration (see Ref. [31] for their most recent global PDF fit) have parametrised PDFs using neural networks, which shall be the most relevant parametrisation in this thesis. The initial-scale form of the PDFs adopted by NNPDF is given by:

$$xf(x, Q_0^2) = x^\alpha (1-x)^\beta \mathrm{NN}(x; \mathbf{w}), \tag{1.101}$$

where $\mathrm{NN}(x; \mathbf{w})$ is the output of a neural network parametrised by $\mathbf{w}$ (the details of the architecture, and how it is extended appropriately for this thesis, are given in Sect. 4.3.1), and α, β are fixed prior to the fit;[19] the fit is then iterated to check stability of α, β. The main advantage claimed by NNPDF in using a neural network parametrisation is that it allows the exploration of an extremely large space of functions, removing the bias inherent in fixed functional form approaches. This is supported by so-called 'universal approximation theorems' (see Ref. [34], for example), which tell us that any given function can be well-approximated by a sufficiently deep neural network.

1.3.2 The loss function

Once we have decided on a functional form with which to model the PDFs, we must obtain the parameters in the model by the minimisation of a loss function on the global dataset. The natural first-choice for a loss function is the χ^2-statistic, defined by:

$$\chi^2(\mathbf{w}) = (\mathbf{d} - \mathbf{t}(\mathbf{w}))^T \Sigma^{-1} (\mathbf{d} - \mathbf{t}(\mathbf{w})), \tag{1.102}$$

where $\mathbf{d}$ is the vector of experimental central values, $\mathbf{t}(\mathbf{w})$ is the vector of corresponding theory predictions for the experimental datapoints (dependent on the model parameters $\mathbf{w}$), and Σ is the experimental covariance matrix describing correlations between the data. The intuition for this choice of loss function is that we wish the theory predictions to be close to the data, but if the data is more uncertain, we should not require the agreement between data and theory to be as precise. This can be most easily seen in the case that the data is uncorrelated, in which case the covariance matrix is diagonal, and the χ^2-statistic reduces to:

$$\chi^2(\mathbf{w}) = \sum_{i=1}^{N_{\mathrm{dat}}} \frac{(d_i - t_i(\mathbf{w}))^2}{\sigma_i^2}, \tag{1.103}$$

[19] Technically fixed for each *replica* in the fit individually; see the below discussion of the Monte Carlo replica method of error propagation used by the NNPDF collaboration. On a separate note, it has recently been shown in Ref. [33] that the scaling prefactor is not in fact necessary, and the network can learn this behaviour.

where N_{dat} is the total number of datapoints, and σ_i is the uncertainty on the ith datapoint; when the ith datapoint is very uncertain, i.e. σ_i is large, we have that the contribution to the χ^2-statistic from the difference between data and theory at the ith datapoint is suppressed. The form Eq. (3.27) retains this interpretation in the case that the data is additionally correlated.

Positivity. It is necessary to make some modifications to this naïve loss function in modern PDF fits. Firstly, the requirement that PDFs are positive, described in Sect. 1.2 above, can be encoded into the loss function through the use of penalty terms:[20]

$$\chi^2_{\text{pos}}(\mathbf{w}) = \chi^2(\mathbf{w}) + \sum_q \Lambda_q \sum_{k=1}^{N_{\text{grid}}} \text{ELU}_\alpha \left(-f_q(x_k, 5\text{ GeV}^2)\right), \tag{1.104}$$

as described in Eq. (3.10) of Ref. [31]. The sum is over all fitted PDF flavours q, and $x_1, x_2, ..., x_{N_{\text{grid}}}$ is the x-grid on which we demand positivity. The scale $Q^2 = 5$ GeV^2 is chosen to be above the charm mass; this is for technical reasons explained in Ref. [26], wherein it is shown that massive quark PDFs need not be positive below their mass threshold. DGLAP evolution preserves positivity beyond the initial chosen positivity scale.

The function ELU_α is the *exponential linear unit*, defined by:

$$\text{ELU}_\alpha(t) = \begin{cases} t, & \text{if } t \geq 0; \\ \alpha(e^t - 1), & \text{if } t < 0, \end{cases} \tag{1.105}$$

which is intended to penalise negative PDFs but pass positive PDFs (of course other choices of functions would work for this purpose too). The parameters Λ_q and α are hyperparameters, which are chosen to maximise fit quality (indeed, in the NNPDF framework, they are chosen by a hyperoptimisation procedure, described in detail in Sect. 3.3 of Ref. [31] - additionally, the parameters Λ_q are increased at each step of the minimisation to improve convergence).

The t_0 prescription. The second modification to the loss function comes from a slightly unexpected place: we must modify the loss function when we have datasets that include multiplicative normalisation uncertainties. A full discussion of why this is the case is given in Ref. [35], but here we give some brief intuition.

Consider a fit which includes two measurements of the same observable, say d_1, d_2 with

[20] In fact, in the NNPDF fits, further penalty terms are also included to ensure that certain benchmark pseudo-observables (called *positivity datasets*) are positive. This is required because positivity of the PDFs does not guarantee positivity of cross-sections, and vice-versa.

Additionally, NNPDF also impose similar penalties ensuring the *integrability* of the PDFs, but we omit the details as the idea is essentially the same.

$d_1 < d_2$. Suppose further that each of these measurements carries an equal multiplicative normalisation uncertainty, s; in particular, this implies that the absolute error on d_1 is sd_1 and the absolute error on d_2 is sd_2. Now, despite the experimentalists reporting the same error for both datapoints, when minimising the naïve χ^2, there will be a bias towards the value d_1 since it carries the smaller uncertainty. This bias is called the *d'Agostini bias*, after the author who first introduced it in Ref. [36]; minimal examples with far greater mathematical detail are presented in Sect. 3.1 of Ref. [35].

The bias can be countered using the so-called t_0 *prescription* for the χ^2-statistic, as described in Ref. [35]. In brief, this involves the replacement of the usual experimental covariance matrix by a covariance matrix of theory predictions, modifying it to the so-called t_0 covariance matrix, Σ_{t_0}. The matrix of theory predictions that we compute depends on a PDF set, called the t_0 *PDF set*; however, the dependence on this set is usually weak, and in the NNPDF methodology, fits are *iterated* (that is, a new fit is run using the output of the initial fit as the t_0 PDF set) to check stability.

1.3.3 Minimisation of the loss function

To actually perform a PDF fit, we must minimise the chosen loss function through some optimisation method. This is a highly non-trivial task, since the loss function as a function of $\mathbf{w}$ is typically extremely complicated, with multiple local (possibly degenerate) minima. Various algorithms exist to perform the minimisation; here we shall describe only the standard method used by the NNPDF collaboration, namely *stochastic gradient descent*.[21] See Ref. [38] for the first discussion of the use of gradient descent methods in the NNPDF framework.

Gradient descent. Before describing stochastic gradient descent, it is useful to describe *deterministic* gradient descent first. The algorithm works as follows. Suppose that we wish to determine a local minimum of the loss function $\chi^2_{t_0+\text{pos}}(\mathbf{w})$ as a function of the PDF parameters.

(1) Choose some initial values of the PDF parameters, $\mathbf{w} = \mathbf{w}_0$. Choose also some value for the *learning rate*, γ, which should be a positive real number.

(2) Given $\mathbf{w}_n$ for $n = 0, 1, ...$, define $\mathbf{w}_{n+1} = \mathbf{w}_n - \gamma \nabla \chi^2_{t_0+\text{pos}}(\mathbf{w}_n)$. Repeat until convergence is sufficient.

Intuitively, we should expect the algorithm to successfully find a local minimum because $\nabla \chi^2_{t_0+\text{pos}}$ is the direction of fastest increase of the function $\chi^2_{t_0+\text{pos}}$. Therefore, at each step of the algorithm (2), we move in a direction proportional to the direction of fastest

[21] Although previously NNPDF minimised the loss function by use of a *genetic algorithm*, see Ref. [37].

decrease of the function $\chi^2_{t_0+\text{pos}}$. As we approach a local minimum, $\nabla\chi^2_{t_0+\text{pos}}$ becomes shallower and shallower, until we converge to the minimum itself.

The *learning rate* γ determines the speed with which we approach the minimum. This is a hyperparameter which should be chosen to result in a good fit. Choosing too large a value of γ will result in too large step sizes initially, essentially resulting in a random walk around the space. Choosing too small a value of γ will result in too small step sizes initially, and convergence will take a very long time. It is also common to vary the learning rate, decreasing it systematically as the steps of the algorithm proceed and we require greater precision approaching the minimum.

Stochastic gradient descent. Stochastic gradient descent is a slightly more efficient version of gradient descent, since it avoids computing the gradient $\nabla\chi^2_{t_0+\text{pos}}$ in its entirety, which can be costly in a global fit with many datasets. Instead, the dataset is split into a number of smaller *batches*, and at each step (2) of the gradient descent algorithm presented above, the gradient is computed for the loss *only* on one of the smaller batches, chosen at random. This can significantly increase the speed of the fits, and can be shown not to compromise the accuracy.

Cross-validation. Naïve minimisation of the loss function via stochastic gradient descent may lead to overfitting. To combat this, one can use standard cross-validation techniques. In particular, we can split the dataset into a 'training' set and 'validation' set; we then only include the training data in the loss, and monitor the validation statistic whilst training is performed using stochastic gradient descent. At the point at which the loss computed on the validation increases, we know we have reached a point of overfitting. This is the method currently used by NNPDF in their most recent analysis [31].[22]

1.3.4 Error propagation

Performing a fit of PDF parameters using the loss function described above will only lead to a best-fit PDF. Naturally, it is important to also give an estimate of the error on the fit, propagated from the error on the experimental data. Fitting collaborations use several methods to achieve this; here, we shall describe only the *Monte Carlo replica method*, which is the standard method used by the NNPDF collaboration (first introduced in [39] in the context of PDF fits), and will be the only method considered in this thesis.

Suppose we are given a vector of experimental central data $\mathbf{d}$ with corresponding covariance matrix Σ. The Monte Carlo replica method begins by generating an ensemble

[22]Technically, the training-validation split is performed on a replica by replica basis; see Sect. 1.3.4 below.

of *pseudodata replicas*, $\mathbf{d}_1, ..., \mathbf{d}_{N_{\text{rep}}}$, drawn from the multivariate normal distribution:

$$\mathbf{d}_p \sim N(\mathbf{d}, \Sigma). \tag{1.106}$$

For each pseudodata replica, we now perform a PDF fit. The loss function in the ith PDF is defined to be:

$$\chi^2_{i,t_0+\text{pos}}(\mathbf{w}) := \chi^2_{i,t_0}(\mathbf{w}) + \sum_q \Lambda_q \sum_{k=1}^{N_{\text{grid}}} \text{ELU}_\alpha \left(-f_q(x_k, 5\ \text{GeV}^2)\right), \tag{1.107}$$

where

$$\chi^2_{i,t_0}(\mathbf{w}) := (\mathbf{d}_i - \mathbf{t}(\mathbf{w}))^T \Sigma^{-1}_{t_0} (\mathbf{d}_i - \mathbf{t}(\mathbf{w})). \tag{1.108}$$

That is, we use the χ^2-statistic (in the t_0-prescription, with the positivity penalty term) evaluated on the *pseudodata* rather than the *experimental central data*. The resulting PDF is called the ith *PDF replica*, and is the best-fit PDF on the ith pseudodata replica.

The result is an ensemble of PDF replicas, $f_1, ..., f_{N_{\text{rep}}}$, the spread of which are expected to give an indication of the uncertainty on the PDF fit. In particular, statistical estimators are usually calculated from this ensemble, namely the *central* or *mean* PDF:

$$\langle f \rangle := \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} f_i, \tag{1.109}$$

and the *variance* on the PDF:

$$(\Delta f)^2 := \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} f_i^2 - \langle f \rangle^2 . \tag{1.110}$$

Confidence intervals can also be calculated, based either on taking appropriate quantiles of the Monte Carlo replicas, or by computing standard deviations from the central PDF (the latter assumes a Gaussian distribution of the PDFs at each point in x-space).

Before proceeding, it is worth noting that despite the very intuitive nature of the Monte Carlo replica method, the method has significant shortcomings that can sometimes result in unfaithful errors (see in particular App. E of [40]). These issues will be discussed in detail in Chapters 4 and 6; indeed, the problems with Monte Carlo error propagation may prove consequential in future PDF fits.

1.4 Global fits of PDFs and theory parameters

So far, we have focussed exclusively on fitting the finite set of parameters which describe initial-scale PDFs, once we have assumed a fixed functional form. However, PDFs are not the only quantities which enter theory predictions for collider experiments. For example, we saw in Section 1.1 that the DIS structure function F_2 is described at next-to-leading-order in QCD as:

$$\begin{aligned} F_2^{\text{LO+NLO}} = x\sum_q e_q^2 \int_x^1 \frac{du}{u}\, f_q^{\overline{\text{MS}}}\left(\frac{x}{u}, \mu_F^2\right) \Bigg(\delta(1-u) + \frac{\alpha_S}{2\pi} P_{qq}(u) \log\left(\frac{Q^2}{\mu_F^2}\right) \\ + \frac{\alpha_S C_F}{2\pi}\left[\frac{1+u^2}{1-u}\left(\frac{\log(1-u)}{u} - \frac{3}{4}\right) + \frac{5u+9}{4}\right]_+ \Bigg) \\ + x\sum_q e_q^2 \int_x^1 \frac{du}{u}\, f_g^{\overline{\text{MS}}}\left(\frac{x}{u}, \mu_F^2\right) \frac{\alpha_S}{2\pi} \Bigg(P_{qg}(u) \log\left(\frac{Q^2}{\mu_F^2}\right) \\ + T_R\left[(u^2 + (1-u)^2)\log\left(\frac{1-u}{u}\right) - 1 + 8u(1-u)\right]\Bigg). \end{aligned} \tag{1.111}$$

In particular, it carries a dependence on *both* the PDFs *and* the strong coupling α_S. If we were to use data on the structure function F_2 to extract the PDFs only, we would necessarily fix α_S to a particular choice.[23] Thus our PDF set would be a PDF set produced *under the assumption* that the strong coupling takes a given, fixed value. If we were to use this PDF set to make predictions for some new process, for consistency we would need to take the *same* value for α_S in the corresponding hard cross-section for the process.

This can become problematic when we wish to *fit* the strong coupling. For suppose that we wish to extract the value of α_S from some new experimental process, which involves initial-state hadrons. If we make predictions for this fit using PDFs that have been produced under the assumption $\alpha_S = \alpha_{S0}$, for some fixed value α_{S0}, then we risk introducing a bias towards this value in the subsequent determination of α_S from the new data. This is described in significantly more detail, for the specific problem of α_S extractions, in Ref. [41].

Importantly, the above argument applies equally well to *any* theory parameters, not just the strong coupling. Previous work has studied the simultaneous extraction of PDFs and SM parameters; for example, see Ref. [42] for a simultaneous determination of PDFs and α_S, using the 'correlated replica' method.

On the other hand, in this thesis, we shall concern ourselves with theory parameters

[23] More precisely, the value of α_S at some choice of renormalisation scale, for example $\alpha_S(m_Z^2)$, i.e. the value of α_S at the mass of the Z-boson squared.

drawn from *beyond the Standard Model* (BSM) theories (namely *dark matter models* and the *Standard Model Effective Field Theory*). In particular, PDFs are usually fitted under the assumption that the SM is true, which implicitly assumes the additional interaction strengths in any BSM theory are all zero; therefore, before we perform an analysis allowing for BSM physics, using processes with hadrons in the initial state, we should ask ourselves:

- How do the PDFs change between a fit assuming the SM, and a fit where the PDFs are determined simultaneously with the parameters of the BSM theory?
- How do the bounds on the parameters of a BSM theory change between a fit using SM PDFs, compared with a simultaneous fit of the BSM parameters alongside PDFs?

It will be the objective of the first part of this text, *Parton distributions in beyond the Standard Model theories*, to answer these questions in three benchmark scenarios. In Chapter 2, we introduce dark matter models, and describe a toy study of the simultaneous extraction of PDFs together with the mass and coupling of a light, leptophobic *dark photon*. In Chapter 3, we describe the simultaneous extraction of PDFs together with two couplings drawn from the Standard Model Effective Field Theory (SMEFT), using high-energy Drell-Yan data. In Chapter 4, we perform a much more comprehensive joint extraction of PDFs together with all of the SMEFT operators which contribute to processes involving top quarks, using all available LHC Run II top quark data.

In the second part of this text, *Future considerations for fitting parton distributions*, we ask some more general questions about joint PDF-BSM fits. In Chapter 5 we perform a study using 'fake' data which has been generated assuming that the fundamental theory of Nature is in fact the SM plus some New Physics; using this data, we perform PDF fits assuming the SM is in fact *true*, in order to rigorously quantify the error committed. Finally, we conclude in Chapter 6 with a careful analysis of the Monte Carlo replica method used for error propagation in the NNPDF framework (and also as one option in the SMEFiT code, Ref. [43], discussed further in Chapter 4), which we discovered during the course of the study presented in Chapter 4 has significant shortcomings, which may affect future PDF fits.

Part I

Parton distributions in beyond the Standard Model theories

Chapter 2

Parton distributions in a dark matter model

> Stars, everywhere. So many stars that I could not for the life me understand how the sky could contain them all yet be so black.
>
> ---
>
> *from Blindsight,*
> *by Peter Watts*

[This chapter is based on Ref. [44], produced in collaboration with Matthew McCullough and Maria Ubiali. The original idea for the study was Matthew's. The results were produced by myself.]

In Sect. 1.4, it was argued that when performing a fit of parameters in a BSM theory using data which involves initial-state hadrons, one should consider the impact of simultaneously extracting both PDFs and the parameters of the BSM model. In this chapter, we discuss how this issue might be handled in the context of a 'toy' dark matter model, where the New Physics is *light* and *weakly-coupled*.

We begin in Sect. 2.1 with a brief review of dark matter and dark photons, and introduce the 'toy' dark matter model which shall be used in the rest of this chapter. In Sect. 2.2, we describe how PDF evolution is modified in the presence of our proposed dark matter candidate. In Sect. 2.3, we place projected bounds on the mass and coupling of our dark matter candidate by considering the quality of the data-theory agreement on projected high-mass Drell-Yan data. In Sect. 2.4, we discuss how this work might be improved and extended in the future.

2.1 Dark matter and dark photons

Dark matter is the name given to an unknown form of matter, hypothesised as a result of a collection of astronomical observations in the mid 20th century. Some of the first evidence for its existence came from the study of *galactic rotation curves*, which describe the average angular velocity of stars in a galaxy as a function of their distance from the galactic centres. Observational evidence suggests these distributions are flat as we approach the edge of the galaxy; this contradicts the naïve theoretical prediction, based on assuming the mass of the galaxy entirely results from the stars comprising it, which produces an exponential decay of the profile at the edge of the galaxy (see e.g. Fig. 1 of Ref. [45]). Astronomers concluded the existence of additional 'dark' matter in galaxies to explain the unexpected results. Subsequently, evidence for dark matter was supported by cosmological observations. For example, in Ref. [46], it was shown that a universe dominated by baryonic matter alone would result in fluctuations of the Cosmic Microwave Background (CMB) which are not observed in the data; this can be explained by the introduction of an unseen form of weakly-interacting matter, identified once again with dark matter.

In particle physics, we can make progress towards understanding dark matter by hypothesising (non-gravitational) interactions between dark matter and the known SM particles. Proposed dark matter candidates could then be revealed either by *direct detection*, through its production at collider experiments, or via *indirect detection*, by comparing precise theoretical predictions in the SM and dark matter models with equally precise experimental data, and determining which of the two scenarios provides a better explanation of the current data. In this chapter, we shall focus exclusively on indirect detection, which is becoming an increasingly attractive avenue as we enter the high-luminosity phase of the Large Hadron Collider's (HL-LHC) operation; this phase will result in a significant reduction in experimental uncertainties.

Dark photons. The majority of the visible sector cosmological energy budget is comprised of hadrons, yet it is rendered visible by the photon, which itself makes up only a tiny fraction of the energy budget and does not behave as matter. It is not unreasonable to expect that the moment the curtains to the dark sector are drawn back it will be rays of 'dark light' that flood detectors and not necessarily the dominant matter component itself. Thus, the most effective strategy to unveil the particle physics of the dark sector *might* be to search for new light states carrying a vanishingly small fraction of the dark energy budget; perhaps, even, *dark photons* (hereafter referred to as 'B'). In recent years, searches for dark photons have gained momentum, both theoretically and experimentally; see, for example, the recent reviews [47, 48, 49], which paint a picture of the breadth of activities in this area.

Being Abelian vectors, dark photons can naturally be light, thus no specific mass scale is particularly deserving of attention than another. As a result experimental search strategies should endeavour to cover as broad a mass range as possible. Below $m_B \lesssim 1$ GeV a variety of intensity frontier experiments have significant sensitivity to the presence of dark photons, however above this mass scale only high energy accelerators have the capability to probe dark photon parameters.

Pursuing this program, [50, 51, 52] use deep inelastic scattering (DIS) data from HERA, and projected data at the upcoming Electron-Ion Collider (EIC) and Large Hadron Electron Collider (LHeC), to derive bounds on a particular class of dark photon models, in which the dark photon is introduced via kinetic mixing with the SM electroweak bosons. In these studies, the dark photon is treated as a mediator of DIS, hence modifying the theoretical expressions for the DIS structure functions, which allows for the extraction of bounds. Further, as we noted in Sect. 1.4 generally, and is also noted in [51], a fully-consistent treatment using this approach requires a simultaneous fit of both parton distribution functions (PDFs) and dark photon parameters; here, the interplay is a mild second-order effect, yielding a small relaxation of the constraints derived in [50] (however, as we shall see in Chapter 3, at the reach and precision of the high-luminosity phase of the Large Hadron Collider, simultaneous analysis of PDFs and BSM effects will be significantly more impactful).

What if a dark photon was baryonic, being primarily coupled to quarks in preference to leptons? In this case, PDF effects take centre-stage, and it becomes reasonable to consider the dark photon not simply as a mediator of DIS, but as a *constituent* of the proton in its own right.

This is not without precedent; whilst the vast majority of New Physics (NP) searches at the LHC involve processes initiated by coloured partons, namely quarks and gluons, it is well-known that quantum fluctuations can give rise to non-coloured partons inside hadrons, although with much smaller abundance. A key example is the inclusion of photons and leptons as constituents of the proton, which can play a crucial role in achieving precise phenomenological predictions at the LHC. In the recent `LUXqed` publication it was shown that the photon PDF can be determined in a model-independent manner, using DIS structure function data [53, 54]. These results brought an extremely accurate determination of the photon PDF, that superseded the previous model-driven or purely data-driven analyses [55, 56]; now, the `LUXqed` method has been incorporated in several global PDF sets [57, 58, 59]. Going beyond just photon PDFs, the `LUXqed` approach has since been extended to the computation of W and Z boson PDFs [60], and lepton PDFs [61]. Whilst the impact of the photon PDFs is sizeable in a number of kinematic regions, the impact of lepton PDFs is rather small at Run III. However lepton-initiated

processes will become an important feature in the near future, in particular in the HL-LHC phase, which will provide the largest proportion of new high-energy particle physics data in the next 20 years [62, 63, 64, 65].

In this spirit, we might reasonably ask whether the proton could contain small contributions from a dark photon, the consideration of which could be important in the near future. In this work, we assess the impact of the inclusion in the proton of a new, light baryonic dark photon B with mass in the range $m_B \in [2, 80]$ GeV,[1] coupling primarily to quarks via the effective interaction Lagrangian:

$$\mathcal{L}_{\text{int}} = \frac{1}{3} g_B \sum_q \bar{q} \not{B} q, \tag{2.1}$$

where the dark fine structure constant is of the order $\alpha_B \sim 10^{-3}$. The dark photon's parton distribution enters into the PDF evolution equations in the same way as the photon PDF, except for a flavour-universal coupling and a non-zero mass threshold. The other PDFs, particularly the quarks and antiquarks, are modified by the presence of a dark photon, especially in the large-x region; this gives rise to significantly different predictions for key observables that can be measured at a very high degree of precision at the LHC. In Sect. 2.3, we focus on the high invariant-mass Drell-Yan (DY) differential distributions, whose theoretical description is significantly affected by the distortion of the quark and antiquark PDFs due to the presence of a non-zero dark parton density.

For the first time in the literature, we demonstrate the strong sensitivity to this dark photon's mass and coupling of the precise measurements of the high-mass Drell-Yan tails at the HL-LHC, by looking at the data-theory agreement using the standard PDFs and the PDFs modified by the presence of a non-zero dark photon distribution. Whilst the sensitivity of collider measurements to BSM colored partons in the proton has been shown to be very strong [66, 67] – as one would expect given that light coloured particles very rapidly distort both the DGLAP evolution and the running of α_S [68] – in this chapter, we show that in the near future we will still be able to competitively probe the presence of a dark parton that couples to quarks via a much more subtle QED-like mixing.

2.2 Parton distributions in the dark photon model

In order to produce a PDF set which includes a non-zero dark photon distribution, we follow the method described in [69], which constitutes a first exploration into the effects of the inclusion of lepton PDFs. In that study, simple ansätze for the functional forms of the

[1] The mass range considered here is due to strong constraints from low-energy probes for $m_B < 2$ GeV, and because we wish to treat the theory as *effective*, remaining agnostic to the UV completion - in the region $m_B > 80$ GeV kinetic mixing effects with the Z-boson begin to become important in dark photon models; see Sect. 2.3.1 for further explanation.

light lepton PDFs (electrons and muons) are postulated at the initial PDF parametrisation scale $Q_0 = 1.65$ GeV, based on the assumption that initial-state leptons are primarily generated by photon splitting, while leptons that are heavier than the initial-scale (namely the tau) are dynamically generated at their mass threshold and kinematical mass effects are neglected, as is done for all heavy partons in the Zero-Mass Variable-Flavour-Number scheme (ZM-VFN) [70, 71]. All parton flavours, including the lepton ansätze alongside initial quark, gluon and photon PDFs drawn from some fixed baseline PDF fit, are then evolved using an appropriately modified version of the PDF evolution equations presented in Eq. (1.88), thus producing a final PDF set now including lepton PDFs.

We mirror this method in the study presented in this chapter; in particular, we conjecture an appropriate ansatz for the dark photon distribution at the initial scale (naturally assuming that the dark photon is primarily generated by quark splitting), then evolve this new distribution alongside quark, antiquark, gluon and photon PDFs drawn from a baseline set, using the modified 'dark' DGLAP evolution.[2] Hence, via the interplay between the flavours generated by DGLAP evolution, the resulting quark, gluon and photon PDFs differ relative to the original reference set evolved excluding dark photons, allowing the impact of the dark photon inclusion to be assessed (and, in the subsequent section, bounds from HL-LHC projected pseudodata to be extracted).

In this section, we describe this procedure in more detail. We begin by explicitly showing the modification to the DGLAP equations required by the presence of dark photons in the proton. We then display and discuss the resulting 'dark PDF sets', and compare them to baseline PDF sets excluding the dark photon. In particular, we analyse the dark luminosities, which show an appreciable deviation from their SM counterparts for sufficiently large values of the coupling α_B; this motivates the phenomenological study that we present in Sect. 2.3.

2.2.1 The DGLAP equations in the presence of dark photons

As alluded to in Sect. 1.2, in order to combine QCD and electroweak calculations at hadron colliders, the PDF evolution must be determined using the coupled QCD and QED DGLAP evolution equations [72, 73, 74]. Here, we modify these equations by adding the leading order evolution of a dark photon PDF. In order to assess the impact of such a dark photon PDF in the evolution, it is essential to include all QCD and QED contributions of the same magnitude as the leading dark contribution. Indeed, to include the terms multiplied by $\alpha_B \sim 10^{-3}$ consistently, we must also include the terms multiplied by $\alpha_s \sim 10^{-1}$, $\alpha_s^2 \sim 10^{-2}$, $\alpha_s^3 \sim 10^{-3}$, $\alpha \sim 10^{-2}$ and $\alpha\alpha_s \sim 10^{-3}$ in the evolution (and further it is no loss to include the terms multiplied by $\alpha^2 \sim 10^{-4}$); in particular, we work at NNLO in QCD, NLO in QED, and include QCD-QED interference; furthermore, we always include

[2] We shall shortly mention that the contribution of the lepton PDFs is negligible here; we thus ignore it.

a photon PDF. On the other hand, the lepton PDFs determined in [62, 61, 64, 65] give a contribution that is more than one order of magnitude smaller than the dark photon contributions determined in this chapter, thus we can safely ignore them.

With the orders and flavours specified, the modified DGLAP equations which we use in this work can be stated as:

$$\begin{aligned}
\mu_F^2 \frac{\partial g}{\partial \mu_F^2} &= \sum_{j=1}^{n_f} P_{gq_j} \otimes q_j + \sum_{j=1}^{n_f} P_{g\bar{q}_j} \otimes \bar{q}_j + P_{gg} \otimes g + P_{g\gamma} \otimes \gamma \\
\mu_F^2 \frac{\partial \gamma}{\partial \mu_F^2} &= \sum_{j=1}^{n_f} P_{\gamma q_j} \otimes q_j + \sum_{j=1}^{n_f} P_{\gamma\bar{q}_j} \otimes \bar{q}_j + P_{\gamma g} \otimes g + P_{\gamma\gamma} \otimes \gamma \\
\mu_F^2 \frac{\partial q_i}{\partial \mu_F^2} &= \sum_{j=1}^{n_f} P_{q_i q_j} \otimes q_j + \sum_{j=1}^{n_f} P_{q_i\bar{q}_j} \otimes \bar{q}_j + P_{q_i g} \otimes g + P_{q_i\gamma} \otimes \gamma + P_{q_i B} \otimes B \\
\mu_F^2 \frac{\partial \overline{q}_i}{\partial \mu_F^2} &= \sum_{j=1}^{n_f} P_{\bar{q}_i q_j} \otimes q_j + \sum_{j=1}^{n_f} P_{\bar{q}_i\bar{q}_j} \otimes \bar{q}_j + P_{\bar{q}_i g} \otimes g + P_{\bar{q}_i\gamma} \otimes \gamma + P_{\bar{q}_i B} \otimes B \\
\mu_F^2 \frac{\partial B}{\partial \mu_F^2} &= \sum_{j=1}^{n_f} P_{Bq_j} \otimes q_j + \sum_{j=1}^{n_f} P_{B\bar{q}_j} \otimes \bar{q}_j + P_{BB} \otimes B,
\end{aligned} \tag{2.2}$$

where μ_F^2 is the factorisation scale, n_f the number of active flavours, q_i ($\overline{q}_i$) the parton density of the ith (anti)quark,[3] g the gluon PDF, γ the photon PDF, and B the new dark photon PDF.

As described in Sect. 1.2, the splitting functions are perturbatively calculable order by order in QCD and QED perturbation theory, hence we can decompose the splitting functions into series of the form:

$$\begin{aligned}
P_{ij} = &\left(\frac{\alpha_s}{2\pi}\right) P_{ij}^{(1,0,0)} + \left(\frac{\alpha_s}{2\pi}\right)^2 P_{ij}^{(2,0,0)} + \left(\frac{\alpha_s}{2\pi}\right)^3 P_{ij}^{(3,0,0)} \\
&+ \left(\frac{\alpha}{2\pi}\right) P_{ij}^{(0,1,0)} + \left(\frac{\alpha_s}{2\pi}\right)\left(\frac{\alpha}{2\pi}\right) P_{ij}^{(1,1,0)} + \left(\frac{\alpha}{2\pi}\right)^2 P_{ij}^{(0,2,0)} \\
&+ \left(\frac{\alpha_B}{2\pi}\right) P_{ij}^{(0,0,1)} + \cdots,
\end{aligned} \tag{2.3}$$

where we follow the notation of [22, 23]; the upper indices indicate the (QCD,QED,Dark) order of the calculation (where in this work we have added an additional 'Dark' index, corresponding to the powers of the dark coupling α_B). As described above in Sect. 1.2, the splitting functions $P_{ij}^{(1,0,0)}$, $P_{ij}^{(2,0,0)}$, $P_{ij}^{(3,0,0)}$, $P_{ij}^{(0,1,0)}$, $P_{ij}^{(1,1,0)}$ and $P_{ij}^{(0,2,0)}$ are all known.

The coefficients $P_{ij}^{(0,0,1)}$ can be calculated directly by finding the most collinearly-

[3] We now ease notation; in general, instead of writing f_q for the PDF as we did in the introduction, we simply write it as q.

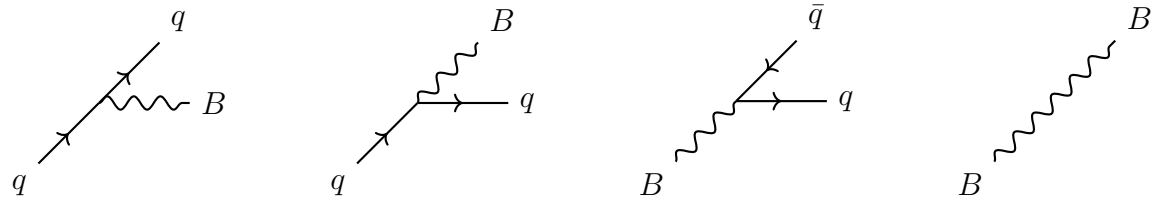


Figure 2.1: The diagrams involving dark photons which contribute to splitting functions. (a), (b), (c), (d) show the contributions to $P_{qq}^{(0,0,1)}(x)$, $P_{Bq}^{(0,0,1)}(x)$, $P_{qB}^{(0,0,1)}(x)$ and $P_{BB}^{(0,0,1)}(x)$, respectively (note at this order, $P_{BB}^{(0,0,1)}(x)$ is proportional to a delta function, $\delta(1-x)$, indicating the lack of possible splitting in this channel).

divergent parts of the four dark splitting channels pictured in Fig. 2.1. The only non-zero contributions are given by $ij = qq, qB, Bq$ and BB (the results are the same for antiquarks). A summary of the calculation is given in App. A of Ref. [44]; however, a detailed calculation is not strictly necessary, since the form of the interaction Lagrangian Eq. (2.1) is identical to that of the electromagnetic-hadronic interaction in the SM, except with a universal coupling $\frac{1}{3}g_B$ to all quarks and antiquarks. It follows that the splitting function contributions provided by the dark photon B will be identical (up to a factor of $\frac{1}{9}$, due to our convention for the universal coupling) to those provided by the photon γ; in particular, we can quote the required leading-order splitting functions by comparing to [69]:

$$\begin{aligned}
P_{qq}^{(0,0,1)}(x) &= \frac{1}{9}P_{qq}^{(0,1,0)}(x) = \frac{1+x^2}{9(1-x)_+} + \frac{1}{6}\delta(1-x), \\
P_{BB}^{(0,0,1)}(x) &= \frac{1}{9}P_{\gamma\gamma}^{(0,1,0)}(x) = -\frac{2}{27}\delta(1-x), \\
P_{qB}^{(0,0,1)}(x) &= \frac{1}{9}P_{q\gamma}^{(0,1,0)}(x) = \frac{x^2+(1-x)^2}{9}, \\
P_{Bq}^{(0,0,1)}(x) &= \frac{1}{9}P_{\gamma q}^{(0,1,0)}(x) = \frac{1}{9}\left(\frac{1+(1-x)^2}{x}\right).
\end{aligned} \tag{2.4}$$

2.2.2 PDF sets with dark photons

We have implemented the modified DGLAP equations described in Sect. 2.2.1 in the public APFEL PDF evolution code [24], which is an accurate and flexible code that can be used to perform PDF evolution up to NNLO in QCD and NLO in QED, using a variety

of heavy flavour schemes. The evolution is performed in x-space,[4] and uses a rotated basis of PDFs such that a maximal number of PDF flavour combinations evolve independently. If we define the following vector of PDFs:

$$\mathbf{q}^S = \begin{pmatrix} g \\ \gamma \\ \Sigma \\ \Delta_\Sigma \\ B \end{pmatrix}, \tag{2.5}$$

where:

$$\Sigma = \sum_{f=u,d,s,c} (f + \bar{f}), \qquad \Delta_\Sigma = \sum_{f=u,c} (f + \bar{f}) - \sum_{f=d,s} (f + \bar{f}), \tag{2.6}$$

then we can choose further independent flavour combinations of PDFs, spanning the complete space of PDFs, such that all of the remaining flavour combinations' evolution equations decouple; this greatly simplifies the computational work. The remaining matrix equation for $\mathbf{q}^S$ can be shown to take the form:

$$\mu_F^2 \frac{\partial \mathbf{q}^S}{\partial \mu_F^2} = \left(\begin{array}{cccc|c} & & & & 0 \\ & \ddots & \ddots & & 0 \\ & \ddots & \ddots & & P_{qB} \\ & & & & P_{qB} \\ \hline 0 & 0 & P_{Bq} & 0 & P_{BB} \end{array}\right) \otimes \mathbf{q}^S. \tag{2.7}$$

Here, the dots denote the relevant SM matrix, with the quark-quark splitting function corrected with a dark contribution as appropriate. This equation (together with the other decoupled scalar equations) is solved using an adaptive step-size fifth-order Runge-Kutta method, as described in [24].

To solve the modified DGLAP equations (2.2), we must also specify initial conditions for the dark photon; this is where we make appropriate ansätze for the functional form of the dark photon at the initial scale $Q_0 = 1.65$ GeV. If the mass of the dark photon m_B were less than the scale Q_0, we could postulate a functional form for the initial dark photon PDF assuming that the dark photon PDF is primarily generated by quark splitting. An appropriate initial condition in this case would be given by:

$$B(x, Q_0^2) = \frac{\alpha_B}{2\pi} \log\left(\frac{Q_0^2}{m_B^2}\right) \sum_{j=1}^{n_f} \left(P_{Bq_j} \otimes q_j(x, Q_0^2) + P_{B\bar{q}_j} \otimes \bar{q}_j(x, Q_0^2) \right). \tag{2.8}$$

[4]Rather than Mellin N-space, which is an alternative obtained by taking the Mellin transform of the DGLAP equations.

On the other hand, our region of interest is $m_B \in [2, 80]$ GeV; in this region, we always have $m_B > 2$ GeV. Thus in our case, we always have $m_B > Q_0$; that is, the mass threshold is always greater than the initial scale. As a result, we set $B(x, Q^2) = 0$ for all $Q < m_B$ and we generate the dark photon PDF dynamically at the threshold $Q = m_B$ from PDF evolution, similar to the treatment of heavy quarks [70, 71], and the tau PDF in [69].

Using the modified code, we produce a PDF set and a corresponding LHAPDF grid [75] including dark photons, for each given value of the dark photon mass and coupling that we consider. We focus on the introduction of a dark photon into the evolution of the `NNPDF3.1luxQED` set [57],[5] which provides our SM baseline, namely an NNLO global PDF analysis of all standard parton flavours together with a photon PDF (the photon PDF in this set is determined using the state-of-the-art `LUXqed` method [54]).

For demonstration purposes, we now proceed to display the key results from a 'dark PDF set' in a particular scenario that is permitted according to the bounds given in Ref. [76], namely:

$$m_B = 50 \text{ GeV}, \qquad \alpha_B = 3 \times 10^{-3}, \tag{2.9}$$

which corresponds to taking $g_B = 1.94 \times 10^{-1}$. As described above, a massless dark photon is generated dynamically at the threshold $Q = m_B$, and is set to zero before this threshold is reached. We have chosen a sufficiently high (admissible) value of the coupling to display the impact upon PDFs and parton luminosities.

In Fig. 2.2, we display both the photon and dark photon PDFs in our representative dark set (obtained by setting the dark photon coupling and mass at the values given in Eq. (2.9)) at the scales $Q = 100$ GeV and $Q = 1$ TeV, and show their relative PDF uncertainties. As anticipated, the dark photon PDF features the same functional form as the photon PDF (this is to be expected since the photon and dark photon splitting functions are identical up to scaling), but its density is smaller since $\alpha_B \lesssim \alpha$. Furthermore, it can be shown that increasing α_B, and also moderately decreasing m_B, increases the similarity of the dark photon and photon PDFs. The dark photon uncertainty is mostly comparable to the photon uncertainty up to $x \sim 0.4$, and then increases faster than the photon uncertainty. This is due to the dark photon being generated off the singlet PDF (the sum of all quarks and antiquarks) at its mass threshold with a rather small coupling; in particular, the dark photon uncertainty is comparable to the uncertainty of the singlet PDF scaled by a factor of α_B. This makes it comparable to the photon PDF uncertainty (for the choice of α_B and m_B of Eq. (2.9)), except in the large-x region where the singlet PDF uncertainty dramatically increases, resulting in the dark photon PDF uncertainty to consistently increase up to $\sim 10\%$ at $x \sim 0.6$. We have verified that for larger couplings

[5]This set will be soon superseded by the PDF set including QED effects obtained starting from NNPDF4.0 [31].

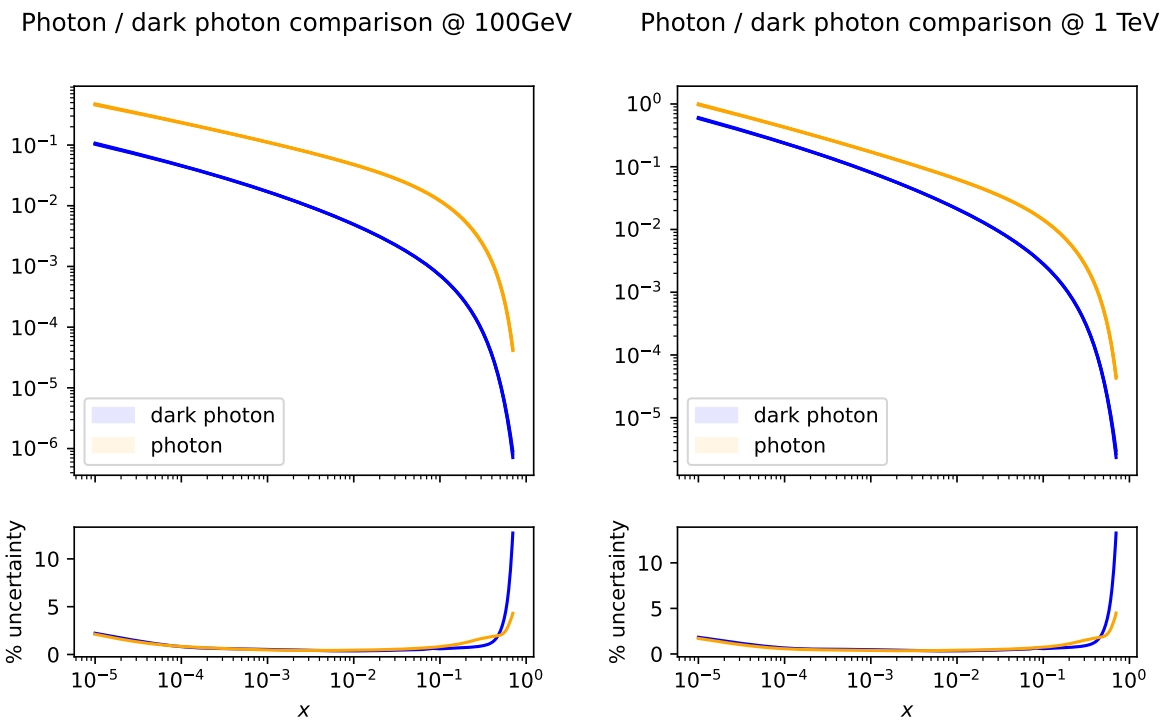


Figure 2.2: Comparison of $x\gamma(x, Q^2)$ and $xB(x, Q^2)$ at $Q = 100$ GeV (left) and $Q = 1$ TeV (right) for the values of dark photon mass and coupling given in Eq. (2.9). The percentage relative 68% C.L. PDF uncertainties of the photon and the dark photon are displayed in the bottom inset.

the uncertainty increases, as one would expect.

Now that we have introduced a new parton in the proton, it is interesting to ask how much 'space' it takes up; this can be quantified by determining the momentum carried by the dark photon at different energy scales. By definition, the momentum fraction carried by any given parton flavour f at the scale Q is given by:

$$\langle x \rangle_f(Q) := \int_0^1 dx\ x f(x, Q). \tag{2.10}$$

In Table. 2.1, we give a comparison between the momentum carried by the dark photon, the photon and the singlet for the representative dark PDF set computed using the values specified in Eq. (2.9), and compare them to the baseline SM PDF set, at $Q = 100$ GeV and $Q = 1$ TeV. We observe that the fraction of the proton momentum carried by the dark photon increases with the scale Q, which is to be expected by analogy with the photon's behaviour. Depending on the coupling and the mass of the dark photon, the latter carries up to a fraction of percent of the proton momentum's fraction at $Q \approx 1$ TeV.

Crucially, the presence of a dark photon in the DGLAP equations also modifies the evolution of all other flavours of PDFs due to the coupling of the PDFs via the modified

$\langle x \rangle_f (Q = 100 \text{ GeV})$	$f = \Sigma$	$f = \gamma$	$f = B$
Baseline	50.23%	0.4241%	0%
Dark set	50.17%	0.4241%	0.03214%
$\langle x \rangle_f (Q = 1 \text{ TeV})$	$f = \Sigma$	$f = \gamma$	$f = B$
Baseline	48.36%	0.5279%	0%
Dark set	48.12%	0.5275%	0.1357%

Table 2.1: A comparison between the momentum fraction percentage carried by the singlet Σ, the photon γ, and the dark photon B at $Q = 100$ GeV and $Q = 1$ TeV, for the baseline SM set and the dark PDF set, obtained with the photon coupling and mass given in Eq. (2.9). The momentum fraction is computed on the central replica in each case.

DGLAP equations Eq. (2.2). We expect that the modification of the quark and antiquark flavours is strongest, as the dark photon is directly coupled to them. We also anticipate a modification to the gluon and photon PDFs, but these will be second order effects, so we expect that they will be smaller in comparison. Moreover, the density of each of the flavours should reduce, as the new dark photon 'takes up space' in the proton which was previously occupied by the quark flavours. Results are shown in Fig. 2.3, in which the ratio between the central value of the dark-photon modified singlet (u-valence) PDF and the central value of the baseline singlet (u-valence) PDF are displayed and compared to the current 68% C.L. PDF uncertainty.

We observe that the modification of the singlet becomes visible at about $x \sim 0.2$ and reaches 3% at larger values of $x \sim 0.5$. This is well within the 68% C.L. uncertainty of the singlet PDF from the baseline `NNPDF3.1luxQED` NNLO set. However, thanks to the inclusion of a vast number of new datasets and the increased precision of the methodology used in global PDF analysis, the recent `NNPDF4.0` NNLO set [31] displays significantly smaller large-x uncertainty. Such a decrease in PDF uncertainties goes in the direction indicated by the dedicated study on how PDF uncertainties will decrease in future, thanks to the inclusion of precise HL-LHC measurements [77]. In particular, to give an indication of how the modification of PDFs due to the presence of a dark photon might come into tension with decreasing PDF uncertainties during the HL-LHC phase, we display the projected 68% PDF uncertainties at the HL-LHC determined in the 'optimistic' scenario, Scenario 3, of Ref. [77]. In this case, should PDF uncertainties decrease to the level predicted by Ref. [77], the distorted singlet PDF approaches the edge of the projected PDF uncertainty at $x \sim 0.1 - 0.3$ region, for the values given in Eq. (2.9). This is particularly relevant for the analysis that we present in the next section.

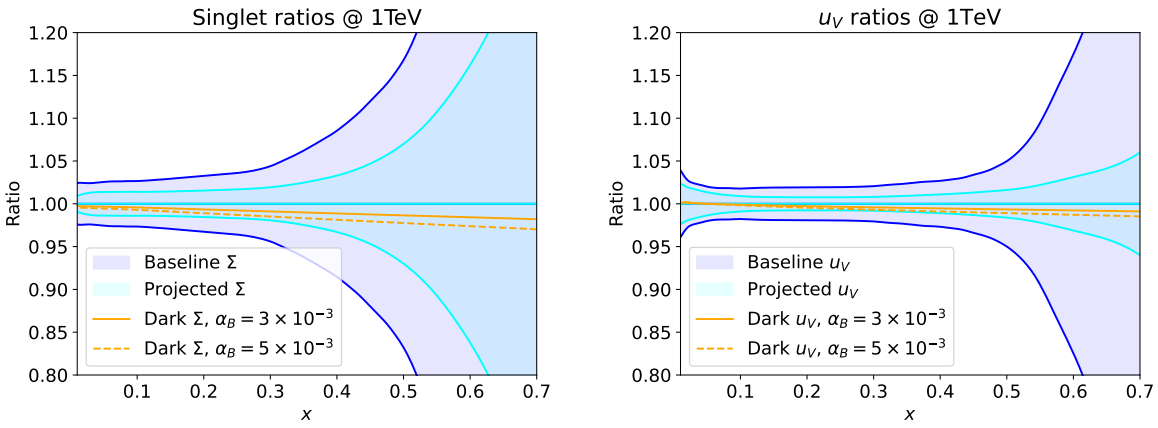


Figure 2.3: In solid orange, the ratio between the central singlet PDF Σ (left) and central u-valence PDF (right), drawn from the dark benchmark scenario in Eq. (2.9), to the baseline SM central replica at $Q = 1$ TeV. In dashed orange, the same ratios but between the SM baseline and a dark PDF set produced using $m_B = 50$ GeV, $\alpha_B = 5 \times 10^{-3}$. In each case, the uncertainty bands represent the 68% C.L. PDF uncertainties of the baseline set (in blue) and the projected PDF uncertainties at the HL-LHC, determined from Ref. [77] (in light blue). The deviation when $\alpha_B = 5 \times 10^{-3}$ approaches the boundary of the projected HL-LHC uncertainty bands, consistent with the behaviour we see in Fig. 2.5 later; increasing α_B (and also to a milder extent, decreasing m_B) pushes the deviation outside of projected HL-LHC uncertainty bands. See the main text for more details.

2.3 Phenomenological implications and projected bounds

In this section we review the existing constraints on the dark photon. Subsequently, in order to assess the impact of a non-zero dark photon parton density on physical observables, we plot the parton luminosities when the dark photon is included, as compared to our baseline SM set. We compare the predicted deviations with the current PDF uncertainties and with the projected PDF uncertainties at the HL-LHC. Finally, we motivate and present an analysis of projected HL-LHC Drell-Yan data and compare the maximal sensitivity we can achieve to the existing bounds derived in the literature.

2.3.1 Review of existing constraints on the dark photon

To appreciate the utility of the dark photon PDF at colliders, we may compare to alternative probes. Recent works considering this class of baryonic dark photon models include [78, 79, 80, 81, 76, 82]. There are a variety of competing constraints on this scenario, of varying theoretical robustness.

One class of constraints, first considered in detail in [78], is theoretical and concerns the mixed $\mathrm{U}(1)_B$−EW anomalies. Suppose we envisage that the UV-completion of the model Eq. (2.1) is perturbative with $\mathrm{U}(1)_B$ linearly realised. In that case, the mixed anomaly must be UV-completed by some fermions with electroweak charges. Early studies

of the classes of fermions that can achieve this include [83, 84].[6] In this perturbative UV-completion they will obtain their mass from spontaneous $U(1)_B$-breaking. As a result, they will be coupled to the longitudinal mode of B and an additional Higgs-like scalar with a Yukawa coupling $\lambda \propto M_F/v_B$, where M_F is the fermion mass, v_B is the $U(1)_B$-breaking expectation value, and we have assumed three sets of fermions with the same charge (1/3) as the left-handed fermions, for simplicity. On the other hand we have $g_B \propto M_B/v_B$ following from the charge and symmetry-breaking vacuum expectation value. As a result, we expect:

$$g_B \approx \frac{2\lambda}{3}\frac{M_B}{M_F}, \tag{2.11}$$

where the precise numerical factors are taken from [78]. Thus, requiring perturbativity $\lambda \lesssim 4\pi$ implies an upper bound on g_B, where the factor 1/3 follows from the fact that each family of fermions is in triplicate to mirror the QCD multiplicity of the SM quarks. This limit is shown as a dashed line in Fig. 2.7 where we have taken $M_F \geq 90$ GeV for the electroweak-charged fermions.

However, a number of implicit assumptions have been made which can weaken upon further inspection. To see this, consider cancelling the anomaly with N copies of the above class of fermions. In this case the limit becomes:

$$g_B \lesssim \frac{8\pi}{3}\frac{N}{3}\frac{M_B}{M_F} \quad . \tag{2.12}$$

Hence we see that this theoretical limit makes not only the assumption of a weakly-coupled UV-completion, but also depends on assumptions of minimality of the UV completion as well. As a result, while this limit does guide the eye as to the nature of the UV-completion, it cannot be considered a strong theoretical limit on the model parameters.

Another constraint which is very relevant in some UV-completions concerns Higgs boson decays. In some UV-completions the radial mode of spontaneous symmetry breaking may mix with the Higgs boson, giving rise to Higgs decays to B's. Depending on the magnitude of the mixing angle the corresponding constraints can be strong, as demonstrated in [88]. Care must be taken to consider these processes in any specific UV-completion, however as the rates depend strongly on the details of the UV-completion we do not include them in our analysis here, which is focussed on the irreducible model-independent IR physics.

The only truly model-independent theoretical limit comes from considering the scale at which the validity of the IR theory itself breaks down. Given that the quark interactions are vector-like there is no possibility of tree-level unitarity violation in quark scattering mediated by B, thus we must look to quantum effects. In this case the mixed-anomaly

[6] Note also that the required fermions could serve as potential dark matter candidates, as discussed in [85, 86, 87].

becomes relevant and renders the theory non-unitary unless [89]:

$$g_B \lesssim \frac{(4\pi)^2}{3\alpha_W}\frac{M_B}{M_\Lambda} \quad , \tag{2.13}$$

where α_W is the SU(2) fine structure constant at the electroweak scale and M_Λ is the energy scale at which the theory becomes strongly coupled. Numerically this is

$$g_B \lesssim \frac{3M_B}{5\text{ GeV}}\frac{10\text{ TeV}}{M_\Lambda} \quad , \tag{2.14}$$

which is too weak to be relevant for our purposes. As a result we conclude that the effective theory considered here is valid throughout the energy scales under investigation. However, we note that, as shown in [82], the mixing with the Z-boson is sensitive to the details of the UV-completion; for this reason we restrict the mass range under investigation to $m_B \leq 80$ GeV, above which these UV-dependent effects can be important.

There are three relevant classes of experimental constraints. The first concerns the exotic Z-boson decays $Z \to B\gamma$. These constraints were calculated in [79] based on the LEP analysis for $Z \to H\gamma$, $H \to$hadrons [90].[7] This limit, relevant to the higher mass range, is shown in red in Fig. 2.7. The second class of constraints at lower masses concerns exotic Υ decays [93, 94], where the constraint is dominated by limits on $\Upsilon(1S) \to 2$ jets [95], shown in blue in Fig. 2.7. Finally, there are additional searches for hadronically decaying resonances at hadron colliders [96, 97, 98, 82]. The strongest are from CMS B+ISR searches [99, 100], shown in yellow in Fig. 2.7.

2.3.2 Effects of the dark photon on parton luminosities

In Sect. 2.2.1, we showed that the presence of a dark photon modifies all other flavours of PDFs via the mixing associated with the DGLAP evolution equations, with a modification that is proportional to α_B and the logarithm of m_B. In order to assess the impact of a dark photon parton density on physical observables, and thus extract the sensitivity that the LHC can achieve on the parameters of the model, in the following subsection we compare the size of the dark parton luminosities to luminosities involving the other partons, and assess the impact of the dark photon on the dominant partonic channels.

Parton luminosities are doubly differential quantities defined as:

$$\frac{d\mathcal{L}_{ij}}{dyd\tau} = f_i(x_1, Q^2)f_j(x_2, Q^2) \qquad x_{1,2} = \sqrt{\tau}\exp(\pm y) \qquad \tau = \frac{M_X^2}{S}, \tag{2.15}$$

where S is the squared centre-of-mass energy of the hadronic collision, M_X is the invariant

[7] Note that this reference does not appear in [79], but instead in [91, 92], however the authors of [79] have confirmed that the limits follow from a recasting of [90].

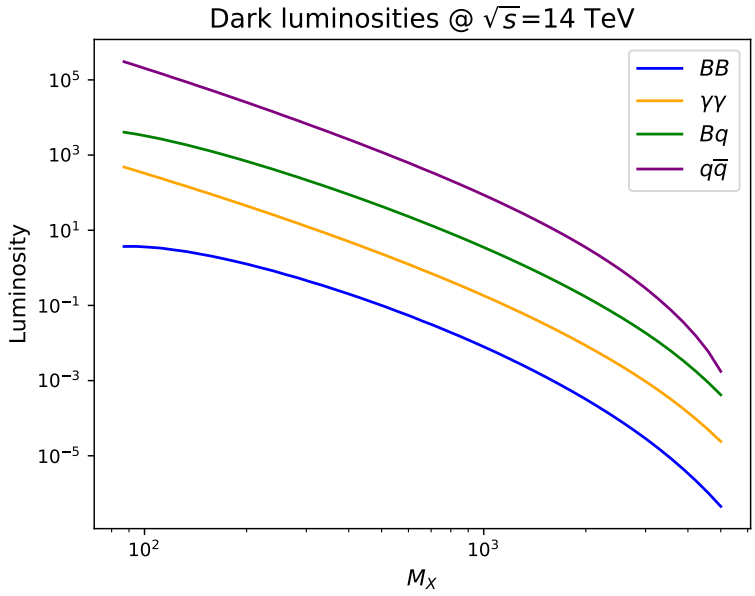


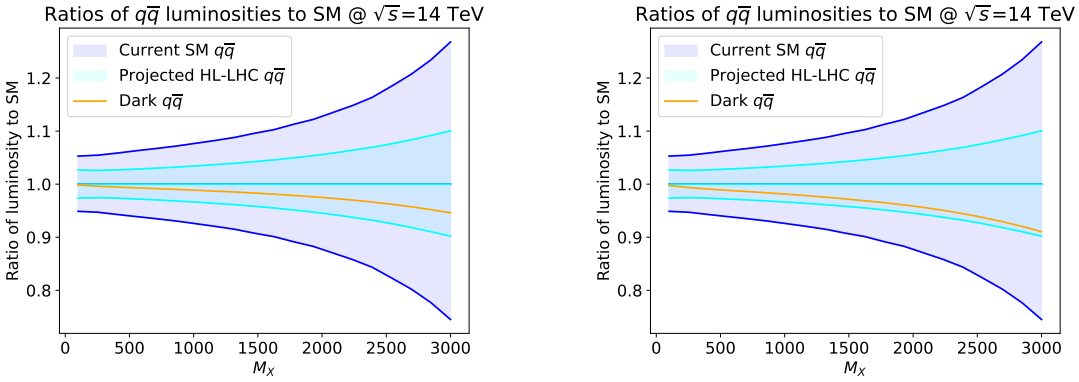
Figure 2.4: Comparison of the absolute value of the Φ_{BB}, Φ_{qB} central luminosities and the $\Phi_{\gamma\gamma}$ and Φ_{qq} central luminosities as a function of the invariant mass M_X at the centre of mass energy $\sqrt{s} = 14$ TeV for the dark PDF set obtained with the dark photon coupling and mass set in Eq. (2.9).

mass of the partonic final state, y is the rapidity of the partonic final state, and $f_i(x, Q^2)$ is the PDF of the ith parton evaluated at the scale Q. Different choices for Q can be adopted in order to improve predictions of a particular process and/or distribution. At the level of pure luminosities, without the convolution with any specific matrix element, the factorisation scale can be naturally set to $Q = M_X$. For plotting purposes, it is useful to define the M_X-differential luminosities, given by:

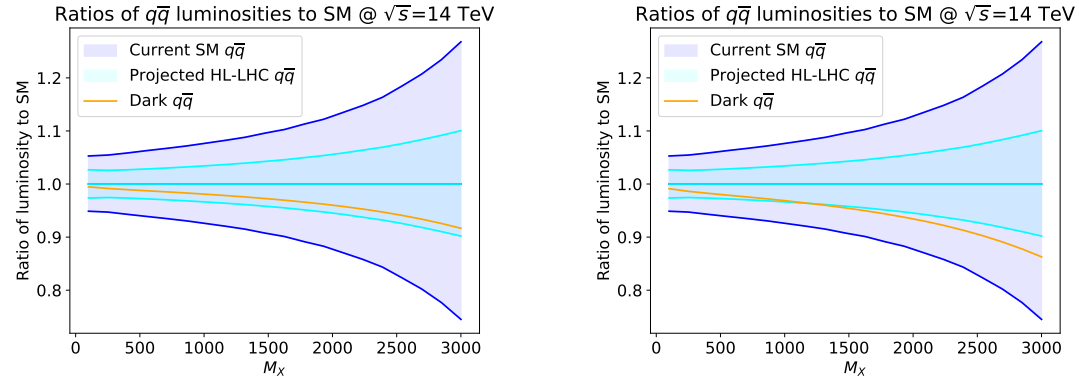
$$\Phi_{ij}(M_X) = \frac{d\mathcal{L}_{ij}}{dM_X^2} = \frac{1}{S} \int_{M_X^2/S}^{1} \frac{dx}{x} f_i(x, M_X^2) f_j\left(\frac{M_X^2}{xS}, M_X^2\right). \tag{2.16}$$

We first compare the size and the M_X-dependence of the different parton luminosities in the candidate dark PDF set obtained by setting the mass and the coupling to the values indicated in Eq. (2.9). In Fig. 2.4 we plot Φ_{BB}, Φ_{Bq} as compared to $\Phi_{q\bar{q}}$, $\Phi_{\gamma\gamma}$. We observe that, while the BB channel is suppressed by two powers of the dark coupling, and its size never exceeds more than a fraction of a percent of the $q\bar{q}$ luminosity, the Bq channel grows from about 2% of the $q\bar{q}$ luminosity at $M_X \sim 1$ TeV to about 8% of the $q\bar{q}$ luminosity at larger values of the invariant mass. Its contribution exceeds that of $\gamma\gamma$ scattering by one order of magnitude.

We now turn to assess the change in the other luminosities, as a result of the inclusion of a non-zero dark photon parton density. In Fig. 2.5 we display the ratio of the dark-



(a) $m_B = 50$ GeV, $\alpha_B = 3 \times 10^{-3}$ (b) $m_B = 50$ GeV, $\alpha_B = 5 \times 10^{-3}$



(c) $m_B = 5$ GeV, $\alpha_B = 3 \times 10^{-3}$ (d) $m_B = 5$ GeV, $\alpha_B = 5 \times 10^{-3}$

Figure 2.5: The ratio $\Phi^{\rm Dark}_{q\bar{q}}/\Phi^{\rm SM}_{q\bar{q}}$ for the total quark-anti-quark luminosity, at the centre of mass energy $\sqrt{S} = 14$ TeV for the values of mass and coupling indicated under each panel. In each panel, the dark blue bands correspond to the current PDF uncertainty, while the light blue bands show the expected uncertainty on the PDF luminosity at the HL-LHC. See main text for more details.

photon modified quark-antiquark integrated luminosity $\Phi^{\rm Dark}_{q\bar{q}}$ with the baseline one, $\Phi^{\rm SM}_{q\bar{q}}$ at the centre of mass energy $\sqrt{S} = 14$ TeV, for different values of the α_B and m_B parameters, starting from our benchmark values, Eq. (2.9). In each figure, the dark blue band corresponds the 68% C.L. PDF uncertainty of the NNLO baseline `NNPDF3.1luxQED` set, while the green bands show the projected PDF uncertainty on the parton luminosity at the HL-LHC; this estimate for the uncertainty on the PDF luminosity is obtained from the 'optimistic' scenario, Scenario 3, analysed in [77], as above. Starting from the values of Eq. (2.9), we observe that the deviation in the $q\bar{q}$ luminosity due to the presence of the dark photon is significant compared to the size of the projected PDF uncertainties at the HL-LHC. Decreasing the mass of the dark photon by a factor of 10 increases the impact of the dark photon on $q\bar{q}$ initiated observables, while increasing the coupling by less than a factor of 2 brings the luminosity beyond the edge of the 68% C.L. error bands.

Crucially, the effect of the dark photon is much larger in the $q\bar{q}$-initiated processes

than in any of the other channels, including qq, qg and gg. This motivates the study of the high-mass Drell-Yan tails that we put forward in the following section.

2.3.3 Constraints from precise measurements of high-energy Drell-Yan tails

Given that the $q\bar{q}$ channel is the most affected by the presence of a non-zero dark photon parton density, in this study we focus on precise measurements of the high-mass Drell-Yan tails at the HL-LHC. It is important to note that these projected data are not included in the fit of the input PDF set used as a baseline, otherwise, as was explicitly shown in [101, 102, 103], the interplay between the fit of the new physics parameters and the fit of the PDF parametrisation at the initial scale might distort the results.

To generate the HL-LHC pseudodata for neutral-current high-mass Drell-Yan cross sections at $\sqrt{S} = 14$ TeV, we follow the procedure of [102]. Namely, we adopt as reference the CMS measurement at 13 TeV [104] based on $\mathcal{L} = 2.8$ fb^{-1}. The dilepton invariant mass distribution $m_{\ell\ell}$ is evaluated using the same selection and acceptance cuts of [104], but now with an extended binning in $m_{\ell\ell}$ to account for the increase in luminosity. We assume equal cuts for electrons and muons, and impose $|\eta_\ell| \leq 2.4$, $p_T^{\rm lead} \geq 20$ GeV, and $p_T^{\rm sublead} \geq 15$ GeV for the two leading charged leptons of the event. We restrict ourselves to events with $m_{\ell\ell}$ greater than 500 GeV, so that the total experimental uncertainty is not limited by our modelling of the expected systematic errors, by making our projections unreliable. To choose the binning, we require that the expected number of events per bin is bigger than 30 to ensure the applicability of Gaussian statistics. Taking into account these considerations, our choice of binning for the $m_{\ell\ell}$ distributions at the HL-LHC both in the muon and electron channels are displayed in Fig. 2.6 with the highest energy bins reaching $m_{\ell\ell} \simeq 4$ TeV. In total, we have two invariant mass distributions of 12 bins each, one in the electron and one in the muon channels.

Concerning uncertainties, in Ref. [102] this data is produced by assuming that the HL-LHC phase will operate with a total integrated luminosity of $\mathcal{L} = 6$ ab^{-1} (from the combination of ATLAS and CMS, which provide $\mathcal{L} = 3\,$ab^{-1} each), and also assuming a five-fold reduction in systematic uncertainty compared to [104]. We regard this scenario as *optimistic* in this chapter; we also manipulated the projected data so that it reflected a more *conservative* possibility, where the total integrated luminosity of the high-mass Drell-Yan tail measurements is $\mathcal{L} = 3$ ab^{-1} (say, for example, they are made available only by either ATLAS or CMS) and with a two-fold (rather than a five-fold) reduction in systematic uncertainties.

For these projections, the reference theory is the SM, with theoretical predictions evaluated at NNLO in QCD including full NLO EW corrections (including in particular

the photon-initiated contributions); note, however, that the Drell-Yan production has been recently computed at N^{3}LO in QCD [105, 106]. In the kinematical region that is explored by our HL-LHC projections ($m_{\ell\ell} > 500$ GeV), the perturbative convergence of the series is good and the N^{3}LO computation is included within the NNLO prediction, with missing higher order uncertainty going from about 1% to a fraction of a percent. Given the good perturbative convergence of the matrix element calculation, and the absence of N^{3}LO PDFs that match the accuracy of the N^{3}LO computation of the matrix element, we use the NNLO QCD and NLO EW accuracy of our calculations, both for the SM baseline and for the dark-photon modified PDF set that we use to compute the maximal sensitivity to the dark photon parameters.

The central PDF set used as an input to generate the theoretical prediction is the SM baseline that we use throughout the paper, namely the NNLO `NNPDF3.1luxQED` set. The percentage statistical and systematic uncertainties on the HL-LHC pseudodata are then estimated as follows. Let us denote by $\sigma_i^{\rm th}$ the theoretical prediction for the DY cross-section from the `luXQED` set, including all relevant selection cuts as well as the leptonic branching fractions. The expected number of events in this bin and the associated (relative) statistical uncertainty $\delta_i^{\rm stat}$ are given by

$$N_i^{\rm th} = \sigma_i^{\rm th} \times \mathcal{L}\,, \quad \delta_i^{\rm stat} \equiv \frac{(\delta N_i)_{\rm stat}}{N_i^{\rm th}} = \frac{1}{\sqrt{N_i^{\rm th}}}\,. \tag{2.17}$$

Note that this bin-by-bin relative statistical uncertainty is the same both at the level of number of events and at the level of fiducial cross sections.

The HL-LHC systematic uncertainties are also estimated from the same reference measurements. If $\delta_{i,j}^{\rm sys}$ denotes the $j^{\rm th}$ relative systematic uncertainty associated to the $i^{\rm th}$ bin of the reference measurement, and if this bin contains $N_i^{\rm th}$ events, then for our projections we assume that the same systematic error associated to a bin with a similar number of expected events will be given by $f_{{\rm red},j}\delta_{i,j}^{\rm sys}$, where $f_{{\rm red},j}$ is the expected reduction in systematic errors foreseen at the HL-LHC (we take the reduction factor to be 0.2 in the *optimistic scenario* and 0.5 in the *conservative scenario*). This assumption is justified since most systematic errors improve with the sample size thanks to *e.g.* better calibration.

Adding in quadrature systematic uncertainties with the statistical error, the total relative uncertainty for the ith bin of our HL-LHC projections is:

$$\delta_{{\rm tot},i}^{\rm exp} = \left(\left(\delta_i^{\rm stat}\right)^2 + \sum_{j=1}^{n_{\rm sys}} \left(f_{{\rm red},j}\delta_{i,j}^{\rm sys}\right)^2 \right)^{1/2}\,, \tag{2.18}$$

where $n_{\rm sys}$ indicates the number of systematic error sources.

The final central values for the HL-LHC pseudodata are then generated by fluctuating

the reference theory prediction by the expected total experimental uncertainty, namely

$$\sigma_i^{\rm hllhc} \equiv \sigma_i^{\rm th}\left(1+\lambda\delta_{\mathcal{L}}^{\rm exp}+r_i\delta_{{\rm tot},i}^{\rm exp}\right)\,, \qquad i=1,\ldots,n_{\rm bin}\,, \tag{2.19}$$

where λ, r_i are univariate Gaussian random numbers, $\delta_{{\rm tot},i}^{\rm exp}$ is the total (relative) experimental uncertainty corresponding to this specific bin (excluding the luminosity and normalisation uncertainties), and $\delta_{\mathcal{L}}^{\rm exp}$ is the luminosity uncertainty, which is fully correlated amongst all the pseudodata bins of the same experiment. We take this luminosity uncertainty to be $\delta_{\mathcal{L}}^{\rm exp}=1.5\%$ for both ATLAS and CMS, as done in [77].

To obtain bounds on the dark photon mass and coupling, we select a grid of benchmark points (m_B, α_B) in the dark photon parameter space; our scan consists of 21 points, distributed as a rectangular grid with masses $m_B = 2, 5, 8, 10, 20, 50, 80$ GeV and couplings $\alpha_B = 10^{-3}, 2\times10^{-3}, 3\times10^{-3}$. We then construct dark PDF sets at each of these benchmark points (thus a total of 21 PDF sets, in each case including quarks, antiquarks, the gluon, the photon and the dark photon PDFs), using the appropriate values of m_B, α_B, and hence compute theoretical predictions in both the *optimistic* and *conservative* scenarios at each grid point. The predictions are produced assuming that the primary contribution comes from the $q\bar{q}$ channel; in particular, we note that the partonic diagrams that include a dark photon in the initial state (such as $Bq \to \bar{q}l^+l^-$ or $B\bar{q} \to ql^+l^-$) are suppressed by two powers of α_B, one from the dark photon PDF and one from the matrix element, and therefore are suppressed beyond the accuracy of our calculation.

In Fig. 2.6 we display the data-theory comparison between the HL-LHC pseudodata in the electron channel, generated according to Eq. (3.45), and both the SM theoretical predictions obtained using the NNLO baseline PDF set `NNPDF3.1luxQED` and the predictions obtained using the dark PDF sets produced with the dark photon mass and coupling set to $(m_B = 5\,\text{GeV}, \alpha_B = 3\times10^{-3})$ and $(m_B = 5\,\text{GeV}, \alpha_B = 5\times10^{-3})$ respectively. We also display the ratio between the central values of those predictions and the central values of the pseudodata as compared to their relative experimental uncertainty in both the *optimistic* and *conservative* scenarios. We see that whilst the SM predictions are within the 1σ experimental uncertainty (by construction), the dark-photon modified predictions display significant deviations. In the bottom inset we show the ratio between the predictions obtained in the two representative dark photon scenarios to the central SM theoretical predictions obtained with the the baseline SM PDF set. PDF uncertainties are shown; we display both the current PDF uncertainty of the NNLO baseline PDF set `NNPDF3.1luxQED` and the projected PDF uncertainties at the HL-LHC, obtained as described at the beginning of this section. Comparing the size of the PDF uncertainties to the size of the projected experimental uncertainties at the HL-LHC, we observe that whilst the current PDF uncertainties are comparable to the experimental uncertainties of the projected data, the projected HL-LHC uncertainties are subdominant as compared to

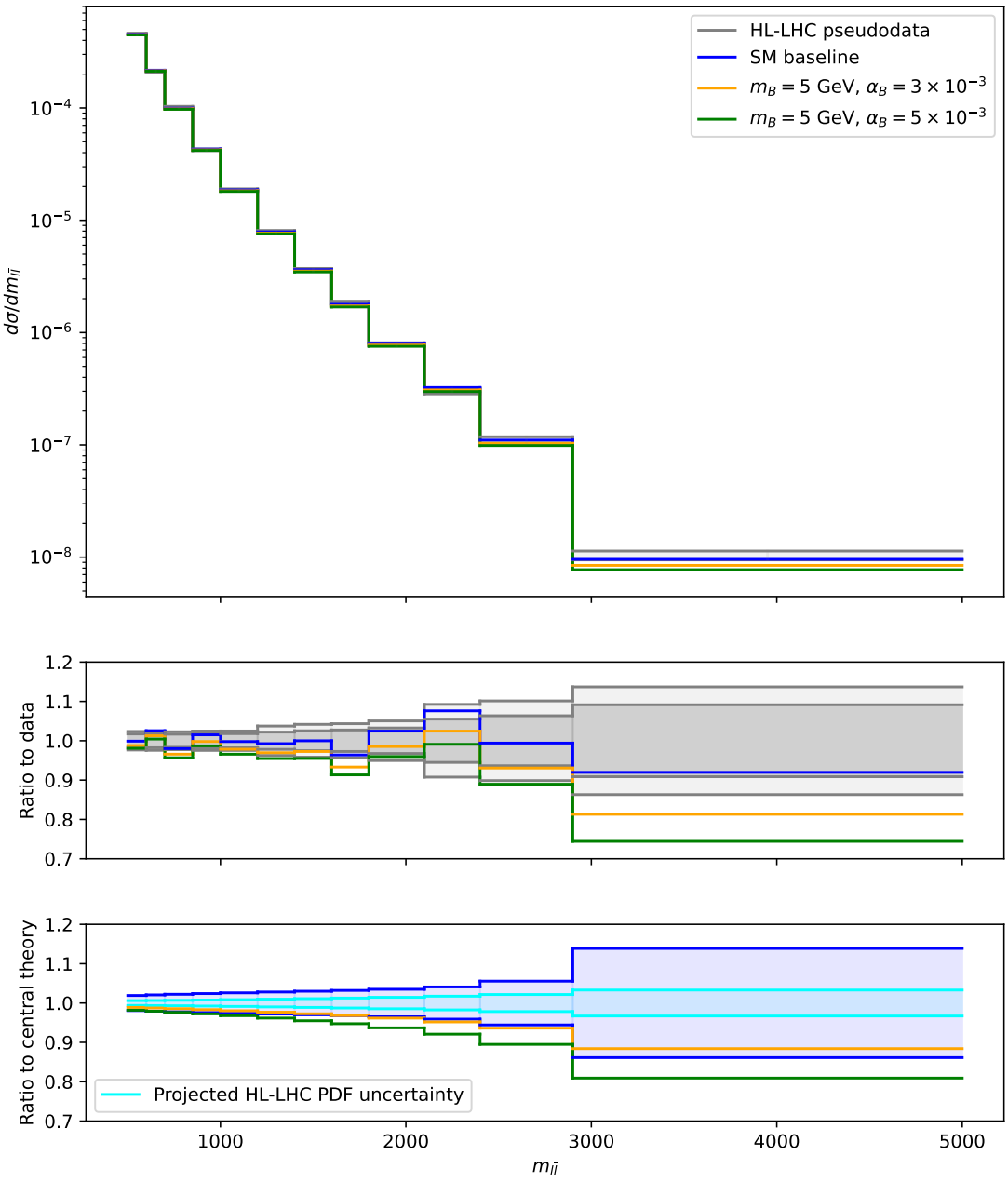


Figure 2.6: **Top:** data-theory comparison between HL-LHC pseudodata in the electron channel generated according to Eq. (3.45) (grey, with *optimistic* uncertainties displayed), and the theoretical predictions obtained using the NNLO baseline PDF set `NNPDF3.1luxQED` (blue) and those obtained using the dark PDF sets produced with parameters $(m_B, \alpha_B) = (5\text{ GeV}, 3 \times 10^{-3}), (5\text{ GeV}, 5 \times 10^{-3})$ (yellow, green respectively). **Middle:** ratio of the baseline SM central predictions obtained using the baseline, and the central predictions obtained using the two representative dark PDF sets, to the central values of the pseudodata. The relative experimental uncertainties in both the *optimistic* scenario (dark grey) and *conservative* scenario (light grey) are displayed. **Bottom:** ratio of the central predictions obtained using the two representative dark PDF sets to the baseline SM central predictions, with both the PDF uncertainty from the baseline PDF set (dark blue) and the projected PDF uncertainty at the HL-LHC in the optimistic scenario of [77] (light blue) displayed.

the experimental uncertainties of the pseudodata.

The χ^2-statistic of the resulting dark PDF set's predictions on high-luminosity high-mass neutral-current DY data is defined as:

$$\chi^2(m_B, \alpha_B) := ||\mathbf{T}(m_B, \alpha_B) - \mathbf{D}||^2_{\Sigma^{-1}}, \tag{2.20}$$

where $||\mathbf{v}||^2_A = \mathbf{v}^T A \mathbf{v}$, $\mathbf{D}$ is the projected data, $\mathbf{T}(m_B, \alpha_B)$ are the theoretical predictions using a dark PDF set containing a dark photon of mass m_B and coupling α_B, and Σ is the total covariance matrix (incorporating both experimental and theoretical uncertainties):

$$\Sigma = \Sigma^{\text{th}} + \Sigma^{\text{exp}}. \tag{2.21}$$

From Fig. 2.6 we observe that, depending on the assumption we make on PDF uncertainties in the HL-LHC era, it may be important to include the PDF uncertainties in the theory covariance matrix, while the component of the theory covariance matrix associated with the scale uncertainty of the NNLO computation is subdominant. Of course, it would be unrealistic to assume that the PDF uncertainty will not decrease as compared to the uncertainty of the `NNPDF3.1luxQED` baseline, given that we already know that in the updated `NNPDF4.0` set [31] the uncertainty of the large-x quarks and antiquarks has already decreased by a sizeable amount thanks to the inclusion of precise LHC data. We thus decide to use the projected PDF uncertainties determined in [77]; in particular, we use Scenario 1 of [77] (the conservative scenario) when we consider the *conservative* experimental scenario, and we use Scenario 3 of [77] (the most optimistic scenario) when we consider the *optimistic* experimental scenario. In Appendix C we discuss how our results depend on the assumptions we make on PDF uncertainties. Assuming that the projected PDF uncertainties at the HL-LHC that we display in the bottom inset of Fig. 2.6 are realistic, even in the most optimistic scenario they still amount to 4% to 6% in the largest bins. Therefore, their contribution is much larger than the scale uncertainty of the Drell-Yan matrix element at NNLO in QCD; hence PDF uncertainty is the dominant theory uncertainty on the predictions, and thus it is this contribution that is included in the theory covariance matrix.

To compute the contribution of PDF uncertainties to the theory covariance matrix, we build the theoretical covariance as defined in [107]:

$$\Sigma^{\text{th}}_{ij} = \langle d\sigma_i^{\text{th},(r)} d\sigma_j^{\text{th},(r)} \rangle_{\text{rep}} - \langle d\sigma_i^{\text{th},(r)} \rangle_{\text{rep}} \langle d\sigma_j^{\text{th},(r)} \rangle_{\text{rep}}, \tag{2.22}$$

where the theoretical predictions for the differential cross section $d\sigma_i^{\text{th},(r)}$ are computed using the SM theory and the r^{th} replica from the baseline PDF set, with PDF uncertainties rescaled by the HL-LHC uncertainty reduction, and averages $\langle \cdot \rangle_{\text{rep}}$ are performed over the

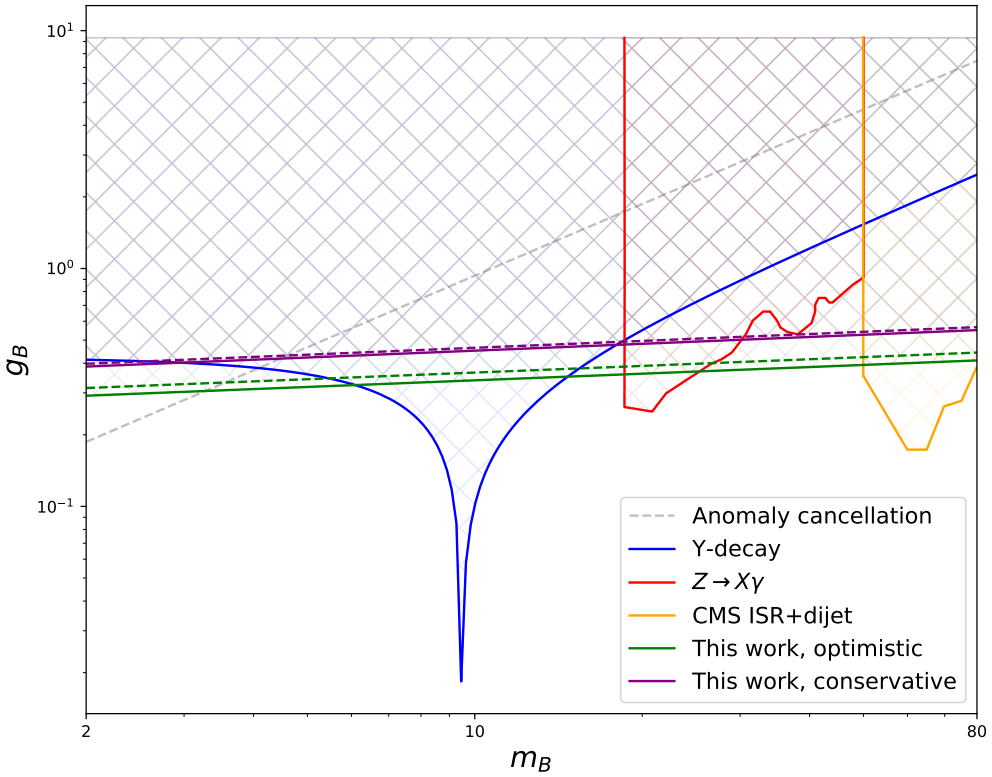


Figure 2.7: Comparison of the projected HL-LHC sensitivity computed in this work in the optimistic (green) and conservative (purple) scenarios with the existing bounds described in Sect. 2.3.1. The solid green and purple lines correspond to projected bounds obtained *excluding* projected PDF uncertainty, whilst the dashed lines correspond to projected bounds obtained *including* projected PDF uncertainty, as discussed in the text.

$N_{\text{rep}} = 100$ replicas of this PDF set.

We define the difference in χ^2 to be:

$$\Delta\chi^2(m_B, \alpha_B) := \chi^2(m_B, \alpha_B) - \chi^2_0, \tag{2.23}$$

where χ^2_0 is the χ^2-statistic when predictions from the baseline set are used instead. For each fixed $m_B = m_B^*$ in the scan, we then model $\Delta\chi^2(m_B^*, \alpha_B)$ as a quadratic in α_B and determine the point at which $\Delta\chi^2 = 3.8$, corresponding to a confidence of 95% in a one-parameter scan. Hence, we construct 95% confidence bounds on m_B, α_B (and hence m_B, g_B via an appropriate conversion) as displayed in Fig. 2.7. There, the purple (dashed) projected bounds are computed in the conservative scenario excluding (including) the PDF theory covariance matrix, and the green (dashed) projected bounds are computed in the optimistic scenario excluding (including) the PDF theory covariance matrix.

We observe that the projected HL-LHC sensitivity to the detection of a dark photon is

competitive with existing experimental bounds, across a large range of possible masses, especially $m_B \in [2, 6] \cup [25, 50]$ GeV. Even in the most conservative scenario including PDF uncertainty (shown as a dashed purple line), the projected sensitivity remains competitive with experiment. Furthermore, one of the useful features of our projected sensitivity is that it is uniformly excluding across a large range (because of the logarithmic dependence on m_B); when compared to individual bounds one at a time, for example the Υ-decay bounds, or the anomaly bounds, our projected sensitivity is powerful.

2.4 Future directions

There are several approximations made in this chapter which should be lifted in future work, and there are various directions that future studies could pursue based on the method presented here.

First, we note that our projected sensitivity is based on the phenomenological tools that we currently have available; we will be able to make more robust statements with the advent of the HL-LHC. Most importantly new global PDF sets will be made available, which will possibly be accurate to $\mathrm{N^3LO}$ in QCD, to be used consistently with $\mathrm{N^3LO}$ computations of the matrix elements. Moreover the associated PDF uncertainty will most certainly include a missing higher order uncertainty component in the PDF uncertainty that is not currently included in any of the global NNLO PDF fits.

Further, the treatment of the dark photon itself could naturally be improved. Here, we make a coarse leading order approximation, based on conjecturing a form for the dark photon PDF and computing the impact of such a PDF on the evolution of the other flavours. Naturally, a more robust analysis would perform a *fit* of a PDF set including dark photon contributions; the expectation would be to treat the dark photon using the `LUXqed` method in the same manner as the photon. There are some technical barriers to such an analysis, but these could be investigated in future works.

Finally, it would be interesting to ask whether similar analyses could be performed for other BSM particles. A particularly interesting candidate is an *axion*, which depending on the model, might couple to gluons and hence have a significant impact on the evolution of the gluon PDF. This could effect interesting processes like top-pair production, extending works such as the recent Ref. [108].

Chapter 3

Parton distributions in the SMEFT from high-energy Drell-Yan tails

[This chapter is based on Ref. [102], produced in collaboration with Admir Greljo, Shayan Iranipour, Zahari Kassabov, Maeve Madigan, Juan Rojo, Maria Ubiali and Cameron Voisey, and additionally on Ref. [109], produced in collaboration with Maeve Madigan. My main contributions to the work comprised: the calculation and production of SMEFT K-factors for all of the DIS processes entering the study, accurate to next-to-leading order in QCD; benchmarking of the charged current DY K-factors against an analytic calculation; running a subset of the fits and analysis, especially in the follow-up study of Ref. [109], presented in Sect. 3.7; collaborative writing of the paper.]

In Chapter 2, we focussed on simultaneous extraction of PDFs and BSM parameters in the case of a dark photon model, where the only BSM particle was *light*. Whilst this model was motivated by the desire to provide a possible dark matter candidate, it is only one of plethora of models which we could have investigated. This suggests the need for a more general framework in which to search for BSM physics.

In the case of *heavy* New Physics, this can be facilitated through the language of *effective field theories* (EFTs). In this chapter, we shall introduce the idea of treating the Standard Model within the effective field theory approach. We shall subsequently present a study wherein we simultaneously extract PDFs and two couplings drawn from the Standard Model EFT, using both DIS and DY data; we shall find that the interplay between PDFs and the Standard Model EFT is mild with the current data, but when we move to using projected HL-LHC data (just as in Sect. 2.3.3), the interplay becomes significant. This study motivates a more comprehensive extraction of PDFs alongside Standard Model EFT couplings in the top sector, which we present in Chapter 4; further, the methodology presented in this chapter is based on a relatively crude *scan* method, whereas the method used in Chapter 4 is much more powerful and flexible. Finally, we

present a first glimpse of extraction of PDFs from projected data *which contains New Physics* in Sect. 3.7, an idea which will be investigated in much greater detail in Chapter 5.

3.1 Effective field theories and the SMEFT

The language of *effective field theory* (EFT) provides a twofold advantage when discussing QFTs: the method of *top-down* EFT can simplify calculations by removing degrees of freedom from a theory which are not relevant at low energy scales, whilst the method of *bottom-up* EFT allows us to parametrise deviations from a good low-energy theory due to unknown high-energy physics.

In this section, we shall provide a brief review of both of these techniques, especially bottom-up EFT as applied to the SM. We begin in Sect. 3.1.1 with a basic demonstration of the key ideas involved in EFT, using a toy model that comprises a single light fermion and a single heavy scalar. In Sect. 3.1.2, we then apply these ideas to the SM, describing how it can be treated in the bottom-up EFT approach resulting in an effective field theory called the *Standard Model Effective Field Theory* (SMEFT). The remainder of the chapter is dedicated to fitting PDFs and two parameters drawn from the SMEFT.

3.1.1 Introduction to effective field theories

To introduce the concept of an *effective* theory, we follow the example of Sect. 4.1 of Ref. [110], and consider a field theory comprising a single real scalar field ϕ and a single fermionic field ψ, with squared masses M^2, m^2 respectively, and with Lagrangian density:

$$\mathcal{L} = \frac{1}{2}(\partial\phi)^2 - \frac{1}{2}M^2\phi^2 + \overline{\psi}(i\partial\!\!\!/ - m)\psi - g\phi\overline{\psi}\psi. \tag{3.1}$$

Suppose additionally that ϕ is much heavier than ψ, with $M^2 \gg m^2$, so that ϕ cannot be produced on-shell in any current experiments, say. The partition function for the theory takes the form:

$$\mathcal{Z} = \int \mathcal{D}\phi\mathcal{D}\psi \exp\left(i\int d^4x \left(\frac{1}{2}(\partial\phi)^2 - \frac{1}{2}M^2\phi^2 + \overline{\psi}(i\partial\!\!\!/ - m)\psi - g\phi\overline{\psi}\psi\right)\right), \tag{3.2}$$

where $\mathcal{D}\phi, \mathcal{D}\psi$ indicate the measures over the relevant functional spaces.[1] By performing the integral over ϕ, we can derive an equivalent theory entirely in terms of ψ; this theory will apply below the scale at which ϕ can be produced on-shell.

To perform the integral, it is convenient to change variables, introducing:

$$\phi'(x) := \phi(x) + ig\int d^4y\ D_F(x-y)\overline{\psi}(y)\psi(y), \tag{3.3}$$

[1]Which of course, in the strictest sense, do not actually exist in a strict mathematical sense.

where D_F is the free-space propagator for the ϕ-field, given by:

$$D_F(z) := \int \frac{d^4k}{(2\pi)^4} \frac{i}{k^2 - M^2 + i\epsilon} e^{-ik\cdot z}. \tag{3.4}$$

Making this change of variables in Eq. (3.2), we obtain:

$$\mathcal{Z} = \int \mathcal{D}\phi' \mathcal{D}\psi \exp\left(i \int d^4x \ \overline{\psi}(i\partial\!\!\!/ - m)\psi\right) \exp\left(i \int d^4x \left(\frac{1}{2}(\partial\phi')^2 - \frac{1}{2}M^2\phi'^2\right)\right)$$
$$\cdot \exp\left(i \int d^4x d^4y \ ig^2 \overline{\psi}(x)\psi(x) D_F(x-y)\overline{\psi}(y)\psi(y)\right). \tag{3.5}$$

The integral over ϕ' can now be performed, producing an irrelevant overall constant of normalisation. The resulting theory then has the non-local Lagrangian:

$$\mathcal{L}_{\text{eff}} = \overline{\psi}(i\partial\!\!\!/ - m) + ig^2 \int d^4y \ \overline{\psi}(x)\psi(x) D_F(x-y)\overline{\psi}(y)\psi(y). \tag{3.6}$$

In the limit $M^2 \to \infty$, we may expand the denominator of D_F in Eq. (3.4) in powers of k^2/M^2, which to a leading approximation gives:

$$D_F(z) = -\frac{i}{M^2}\delta^4(z), \tag{3.7}$$

hence we may rewrite the Lagrangian Eq. (3.6) as:

$$\mathcal{L}_{\text{eff}} = \overline{\psi}(i\partial\!\!\!/ - m)\psi + \frac{g^2}{M^2}(\overline{\psi}\psi)^2 + O\left(\frac{1}{M^4}\right). \tag{3.8}$$

We say that this theory is an *effective theory*, valid below the energy scale M, with *ultraviolet (UV) completion* given by the original theory. Important features to note are the following:

(i) Integrating out the field ϕ has resulted in a Lagrangian containing a *non-renormalisable* term, $(\overline{\psi}\psi)^2$; this can be seen by counting mass dimensions, which in this case shows that this operator is in fact a dimension *six* operator. In a QFT that we expect to hold to arbitrarily high energy scales, such terms are not permissible. However, in our *effective* theory, there is no such stringent requirement; we indeed expect the theory to break down at the scale M, and so non-renormalisable terms are in fact allowed.

(ii) The non-renormalisable operators in the effective theory have been organised into a series, with their associated couplings suppressed by increasing powers of $1/M^2$. By counting mass dimensions, we see that the higher the dimension of the operator, the more subleading its effect.

Top-down vs bottom-up. The construction above is known as *top-down* construction of an EFT. In this paradigm, we begin with a UV-complete theory, and then we integrate out modes to produce an effective theory containing non-renormalisable operators; the theory is then valid only up to some large energy scale. Top-down construction of an EFT is particularly useful if we wish to simplify calculations by ignoring degrees of freedom which are not relevant at the energy scale of study (for example, the W and Z bosons are often integrated out of the electroweak theory when performing low-energy calculations, resulting in the *Fermi effective theory*).

On the other hand, the EFT approach can also be applied in a *bottom-up* fashion. Suppose that we have a renormalisable theory which works well at low energies, but we expect to break down at some higher energy scale, making way for another theory which contains additional, unknown, heavy degrees of freedom. The low-energy limit of this high-energy theory must match our low-energy theory; in particular, our low-energy theory should be considered the leading approximation in a top-down EFT approach to the unknown high-energy theory. Therefore, we can build more precise approximations to the high-energy theory by adding on all non-renormalisable operators to the low-energy Lagrangian, built from the low-energy degrees of freedom and respecting any symmetries of the low-energy theory that we believe still hold in the high-energy theory. These non-renormalisable operators can be organised in terms of importance by their *mass dimension*, since dimensional analysis tells us their couplings must be increasingly suppressed by a characteristic scale of the high-energy physics (e.g. the mass of the lowest-energy degree of freedom that is integrated out at low energies) as the mass dimension of the operator increases.

For a concrete example, suppose that we believe that at low energies, Nature is described by the fermionic Lagrangian:

$$\mathcal{L}_{\text{low}} = \overline{\psi}(i\not{\partial} - m)\psi. \tag{3.9}$$

This theory has a $U(1)$ global vector symmetry $\psi \mapsto e^{i\theta}\psi$ and a $U(1)$ global axial symmetry $\psi \mapsto e^{i\theta\gamma^5}\psi$, which perhaps we believe to be fundamental (and suppose we also believe that Lorentz symmetry is fundamental). We can describe theories which have the low-energy limit $\mathcal{L}_{\text{low}}$ by adding on all possible higher-dimensional operators which respect the symmetry, organised by their mass dimension, to produce a *bottom-up* EFT which is the low-energy limit of the unknown high-energy theory. Such operators can only be built from $\psi, \overline{\psi}$ and the derivative ∂; the mass dimensions of these objects are $[\psi] = [\overline{\psi}] = 3/2$ and $[\partial] = 1$, respectively. Considering the ways in which these objects can be combined at each mass dimension, we therefore obtain an EFT approximation of the form (up to operators of mass dimension 6):

$$\mathcal{L}_{\text{low}}^{\text{EFT}} = \overline{\psi}(i\not{\partial} - m)\psi - \frac{g_5}{\Lambda}\overline{\psi}\partial^2\psi - \frac{g_6^{(1)}}{\Lambda^2}(\overline{\psi}\psi)^2 - \frac{g_6^{(2)}}{\Lambda^2}\overline{\psi}\partial^2\not{\partial}\psi - ..., \tag{3.10}$$

where Λ is some characteristic scale of New Physics.

Importantly, some of these operators can be removed by considering the *equations of motion* of the theory (a point we shall shortly return to when we discuss EFT bases below). The fermion ψ obeys the free Dirac equation in the low-energy theory, given by:

$$(i\not{\partial} - m)\psi = 0, \tag{3.11}$$

which yields the Klein-Gordon equation $(\partial^2 + m^2)\psi = 0$ when the operator $(i\not{\partial} - m)$ is applied a second time. Therefore, whenever a second derivative term appears, we may replace ∂^2 with $-m^2$. As a result, the effective theory can be simplified to:

$$\mathcal{L}_{\text{low}}^{\text{EFT}} = \overline{\psi}(i\not{\partial} - m)\psi - \frac{g_6}{\Lambda^2}(\overline{\psi}\psi)^2 - \dots \,. \tag{3.12}$$

The reason that we can apply the *free* equations of motion here, even though we believe the fermion is actually coupled in the high-energy theory, is that the corrections to the free equations of motion enter at order $1/\Lambda$, so are suppressed when inserted into the Lagrangian and can be neglected at the order we are considering. The unknown coefficients in the expansion, g_6, etc, are called *Wilson coefficients*.

As expected, if we know that the UV-completion of the theory is given by Eq. (3.1), the bottom-up EFT Eq. (3.12) can be *matched* to the UV-complete theory. In this case, this can be achieved by comparing Eq. (3.8) with Eq. (3.12), where we identify:

$$\frac{g_6}{\Lambda^2} = -\frac{g^2}{M^2}. \tag{3.13}$$

In more complicated scenarios, matching can be performed order-by-order in perturbation theory by computing processes using both the diagrams of the UV-complete theory and the diagrams of the effective theory, and then comparing the results. There are also other methods of matching available, including *functional matching* (see for example Ref. [111]).

Renormalisability. By their very construction, effective theories contain non-renormalisable operators, and hence are not renormalisable theories. However, as we mentioned above, this is no problem in the EFT philosophy, since we do not expect EFTs to hold to arbitrarily high energies.

Further, whilst EFTs are not renormalisable *per se*, they do have the useful property that they are renormalisable *order by order* in $1/\Lambda$, where Λ is the characteristic scale of the New Physics. This manifests in the following way: if we perform a loop calculation in our EFT, we shall find that we obtain divergent contributions which can only be cancelled by counterterms with higher mass dimensions than those considered at a fixed order in $1/\Lambda$. However, the divergent contribution will itself be multiplied by a higher power in

$1/\Lambda$ than the fixed order we are considering; hence, it can be dropped. Details of such a calculation are presented in Sect. 5.4 of Ref. [112].

Renormalisation of an EFT also results in *running* of the Wilson coefficients in the EFT. We will neglect this effect in the rest of the thesis, since it will be subleading in all scenarios considered, but it is worth noting that the coefficients are not constants.

EFT bases. In the case of a more general bottom-up EFT, with many more low-energy degrees of freedom, there will be a vast array of possible operators that can be written down at a given mass dimension in the EFT expansion. Crucially, these operators *may not be independent*. Generically, there are two ways in which operators may be dependent on one another:

(i) **Equations of motion.** The operators might be related by the *equations of motion* of the low-energy theory. We already saw an example of this when reducing Eq. (3.10) to Eq. (3.12).

(ii) **Integration by parts.** The operators might be related by *integration by parts*. Consider, for example, the following pair of operators:

$$\phi^2\partial^2\phi, \qquad\qquad \phi(\partial\phi)^2, \tag{3.14}$$

built from a scalar field ϕ and the derivative ∂_μ. On the surface, these appear to be separate operators. However, they are in fact the same operator, when related by integration by parts:

$$\int d^4x\ \phi^2\partial^2\phi = -\int d^4x\ \partial(\phi^2)\cdot\partial\phi = -2\int d^4x\ \phi(\partial\phi)^2, \tag{3.15}$$

assuming that boundary terms decay at infinity. Therefore, if we were to naïvely insert the sum $g\phi^2\partial^2\phi + g'\phi(\partial\phi)^2$, we would find that it in all processes the Wilson coefficients g, g' would enter in the combination $g' - 2g$; we would never be able to determine them independently of one another.

It follows that in any construction of a bottom-up EFT, to avoid redundancy we should endeavour to find a minimal set of operators (at each mass dimension) which are independent of one another, so that their associated Wilson coefficients can additionally be determined independently of one another. Such a minimal set is called a *basis* for the EFT at the given mass dimension; this notion will be immediately useful when we construct the SMEFT using the so-called *Warsaw basis* in the following section, Sect. 3.1.2.

3.1.2 The SMEFT

The framework of bottom-up EFTs can naturally be applied to the Standard Model itself, producing an EFT known as the *Standard Model Effective Field Theory* (SMEFT) (see e.g. Ref. [113] for a review). Following the procedure outlined above, the Lagrangian of the SMEFT must take the form:

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_{d=5}^{\infty} \sum_{i=1}^{n_d} \frac{c_d^{(i)}}{\Lambda^{d-4}} \mathcal{O}_d^{(i)}, \tag{3.16}$$

where $\mathcal{L}_{\text{SM}}$ denotes the Lagrangian of the Standard Model, and $\mathcal{O}_d^{(i)}$ denotes the ith operator in a basis for the d-dimensional operators built from the SM fields and obeying the SM symmetries (Lorentz symmetry and gauge symmetry). Here, n_d denotes the number of independent operators of mass dimension d, $c_d^{(i)}$ denotes the ith Wilson coefficient at mass dimension d, and Λ represents a characteristic scale of New Physics.

The types of operators that arise at dimension $d = 5$, $d = 6$ in the expansion have been completely classified. The operators at dimension $d = 5$ are known as *Weinberg operators*, and are all lepton number violating; strong constraints on their respective Wilson coefficients from LEP mean that these operators are usually ignored. At dimension $d = 6$, there are a total of $n_6 = 2499$ independent operators, parametrised by the so-called *Warsaw basis*, introduced in Ref. [114]. It is the dimension six terms which we shall be most interested in throughout this thesis. The complete list of operators in the Warsaw basis is presented in Ref. [114] in Tables 2 and 3; we shall only require a small subset of these operators in the sequel.

3.2 Parton distributions in the SMEFT

Accurate measurement of observables at high energies offers one of the most promising avenues towards an indirect discovery of BSM physics at the LHC. While the collider has now (almost) reached the design energy, its integrated luminosity continues to grow steadily, thus greatly facilitating dedicated studies of the (currently statistics-limited) high-energy tails of distributions. Furthermore, many higher-dimensional SMEFT operators in the Warsaw basis introduced in Sect. 3.1.2 induce deviations from the SM which grow with the energy of the partonic collision, enhancing the BSM sensitivity of these high-energy tails. For instance, in the Drell-Yan process, the naïve scaling of the $\psi\psi \to \psi\psi$ scattering amplitude in an underlying EFT description leads to an amplitude $\mathcal{A} \propto E^2/\Lambda^2$, where E is the energy of the process and Λ is the New Physics scale. Thus, rather generically, a less precise measurement of a high-mass tail can compete with a low-energy precision measurement due to this energy enhancement, which can be traced back to the preservation

of unitarity. This property leads to the somewhat unintuitive result that the study of the high-energy lepton tails in the Drell-Yan process represents a competitive probe of BSM dynamics compared to electroweak precision tests and low-energy flavour physics test.

However, in order for the SMEFT interpretation of these high-energy tails to convincingly uncover a BSM effect, it becomes crucial to ensure full control over the SM inputs and their uncertainties, such as those associated to the PDFs [115]; as described in Sect. 1.4, this requires us checking that eventual BSM deviations arising in high-energy tails are not being inadvertedly reabsorbed into the PDFs, entailing the need for a simultaneous extraction of PDFs and SMEFT couplings.

A first take on this challenge was presented in a proof-of-concept study [101] where the simultaneous determination of PDFs and four SMEFT coefficients from deep inelastic scattering (DIS) structure functions was demonstrated. There, it was found that the SMEFT corrections can indeed be partially reabsorbed into the PDFs but also that it is possible to robustly disentangle QCD and BSM effects by exploiting their different energy scaling.

The main goal of the rest of this chapter is to extend this approach to LHC processes, specifically with the joint determination of PDFs and EFT coefficients from DIS and Drell-Yan data. Drell-Yan processes in general, and high-mass measurements in particular, provide information on the light quark and anti-quark PDFs in a broad region of x representing an important ingredient in modern global PDF fits [116, 117, 32, 118]. Furthermore, high-mass Drell-Yan data will be instrumental at the High-Luminosity LHC (HL-LHC) to pin down the large-x PDFs [119]. Considering that SMEFT signals can lead to significant deviations from the SM in these same high-energy DY tails, one would like to assess to what extent they can be reabsorbed into the PDFs and to define strategies to separate QCD from BSM effects.

In order to interpret the Drell-Yan data in the SMEFT framework in this chapter, we formulate a simple, yet motivated, benchmark scenario [120]; in particular, we consider the $\hat{W}$ and $\hat{Y}$ electroweak parameters generated in universal theories that modify the electroweak gauge boson propagators and lead to flavour-universal deviations which grow with the invariant mass. This scenario is discussed in detail in Sect. 3.3. Subsequently, in Sect. 3.4 we summarise the datasets (taken to be exclusively DIS and DY sets to ensure a theoretically consistent description of the PDFs in the SMEFT) used in the analysis, and the corresponding theoretical calculations. In Sect. 3.5 we present the results for the simultaneous determination of the SMEFT coefficients and the PDFs from the available high-mass DY data from LHC Run I and Run II, in the two scenarios presented in Sect. 3.3 and assess how they modify the interpretation of BSM searches based on the SM PDFs. In Sect. 3.6 we present a summary of the constraints we find on the two scenarios we consider and we assess the outcome of a joint PDF and SMEFT analysis using projections

for the HL-LHC, just as we did in Sect. 2.3. Finally, in Sect. 3.7, we present the results of a follow-up study where we consider the outcome of a joint PDF and SMEFT analysis using projections for the HL-LHC which have had New Physics injected into the tails; that is, we ask how our analysis changes *if* New Physics is present during the HL phase of the LHC.

3.3 The SMEFT scenario: oblique corrections

In this section we present the SMEFT benchmark scenario that will be used in this chapter to interpret the LHC Drell-Yan processes. This scenario belongs to the class of electroweak precision tests and is sensitive to a broad range of UV-complete theories proposed in the literature.

The *oblique corrections*, as originally proposed in [121, 122], play a key role in testing theories beyond the Standard Model. They parametrise the self-energy $\Pi_V(q^2)$ of the electroweak gauge bosons W_μ^a and B_μ, where $V = W^3W^3$, BB, W^3B, and W^+W^-. Truncating the momentum expansion at order q^4, while imposing proper normalisation and symmetry constraints, one concludes that there are only four oblique parameters which can be identified with dimension-six operators in the SMEFT. These are the well-known $\hat{S}$, $\hat{T}$, $\hat{W}$, and $\hat{Y}$ parameters [123]. The parameters $\hat{S}$ and $\hat{T}$ are well constrained from precision LEP measurements [123] and grow slowly with q^2, while $\hat{W}$ and $\hat{Y}$ scale faster implying that their effects will be enhanced for the high-energy dilepton tails at the LHC [120]; while $\hat{T} = \mathcal{O}(q^0)$ and $\hat{S} = \mathcal{O}(q^2)$, instead one has that $\hat{W}, \hat{Y} = \mathcal{O}(q^4)$.

Whilst these operators can be written initially as parametrising self-energies of the electroweak bosons,[2] using the equation of motions they can be rotated to be written in terms of the following four-fermion operators:

$$\mathcal{L}_{\text{SMEFT}} \supset -\frac{g^2\hat{W}}{2m_W^2} J_{L\mu}^a J_L^{a\mu} - \frac{g_Y^2\hat{Y}}{2m_W^2} J_{Y\mu} J_Y^\mu \quad , \tag{3.17}$$

where J_L and J_Y are $SU(2)_L$ and $U(1)_Y$ conserved fermionic currents,

$$J_L^{a\mu} = \frac{1}{2}\sum_{f=q,l} \bar{f}\sigma^a\gamma^\mu f \quad , \qquad J_Y^\mu = \sum_{f=q,l,u,d,e} Y_f\, \bar{f}\gamma^\mu f \quad . \tag{3.18}$$

Here q, l are the SM quark and lepton left-handed doublets, while u, d, e are the right-handed singlets. Also, g and g_Y are the corresponding electroweak gauge couplings, σ^a are

[2] See e.g. Ref. [124] where they are written in terms of the *universal basis* for the SMEFT, where they are defined by:

$$\mathcal{L}_{\text{SMEFT}} \supset -\frac{\hat{W}}{4m_W^2}(D_\rho W_{\mu\nu}^a)^2 - \frac{\hat{Y}}{4m_W^2}(\partial_\rho B_{\mu\nu})^2.$$

the Pauli matrices, and the hypercharges are given by $Y_f = 1/6$, $-1/2$, $2/3$, $-1/3$, and -1 for q, l, u, d, e, respectively. Summation over flavour indices is assumed, which implies that in this scenario the fermionic currents respect the $U(3)^5$ global flavour symmetry.

Expanding Eq. (3.17), one can relate the $\hat{W}$ and $\hat{Y}$ parameters to the coefficients of dimension-six operators in the Warsaw basis introduced in Sect. 3.1.2. There, the operators relevant to the description of the Drell-Yan process are given by:

$$\begin{aligned}
&\mathcal{O}_{ld} = (\bar{l}\gamma_\mu l)(\bar{d}\gamma^\mu d)\,, \qquad \mathcal{O}_{lu} = (\bar{l}\gamma_\mu l)(\bar{u}\gamma^\mu u)\,, \quad \mathcal{O}_{lq}^{(1)} = (\bar{l}\gamma^\mu l)(\bar{q}\gamma_\mu q)\,,\\
&\mathcal{O}_{ed} = (\bar{e}\gamma_\mu e)(\bar{d}\gamma^\mu d)\,, \qquad \mathcal{O}_{eu} = (\bar{e}\gamma_\mu e)(\bar{u}\gamma^\mu u)\,, \quad \mathcal{O}_{qe} = (\bar{q}\gamma_\mu q)(\bar{e}\gamma^\mu e)\,,\\
&\mathcal{O}_{lq}^{(3)} = (\bar{l}\sigma^a\gamma^\mu l)(\bar{q}\sigma^a\gamma_\mu q)\,.
\end{aligned} \tag{3.19}$$

Note the flavour indices are contracted within the brackets, for example $\bar{l}\gamma_\mu l \equiv \bar{l}^1\gamma_\mu l^1 + \bar{l}^2\gamma_\mu l^2 + \bar{l}^3\gamma_\mu l^3$. Taking into account this matching between the $\hat{W}$ and $\hat{Y}$ parameters and the corresponding Wilson coefficients in the Warsaw basis, we can express the SMEFT Lagrangian in this scenario, Eq. (3.17), as follows:

$$\begin{aligned}
\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} - \frac{g^2\hat{W}}{4m_W^2}\mathcal{O}_{lq}^{(3)} - \frac{g_Y^2\hat{Y}}{m_W^2}\Big(&Y_l Y_d\,\mathcal{O}_{ld} + Y_l Y_u\,\mathcal{O}_{lu}\\
&+ Y_l Y_q\,\mathcal{O}_{lq}^{(1)} + Y_e Y_d\,\mathcal{O}_{ed} + Y_e Y_u\,\mathcal{O}_{eu} + Y_e Y_q\,\mathcal{O}_{qe}\Big)\,.
\end{aligned} \tag{3.20}$$

The parametrisation in Eq. (3.19) was implemented using the `SMEFTsim` package [125] and cross-checked against the reweighting method used in Ref. [126] (see also [127]).

The analysis in Ref. [120] reports the following 95% confidence level intervals on $\hat{W}$ assuming $\hat{Y} = 0$,

$$\begin{aligned}
&\hat{W} \in [-3, 15] \times 10^{-4} \text{ (ATLAS 8 TeV, 20.3 fb}^{-1}\text{ [128])}\,,\\
&\hat{W} \in [-5, 22] \times 10^{-4} \text{ (CMS 8 TeV, 19.7 fb}^{-1}\text{ [129])}\,,
\end{aligned} \tag{3.21}$$

as well as, the 95% confidence level intervals for $\hat{Y}$ assuming $\hat{W} = 0$,

$$\begin{aligned}
&\hat{Y} \in [-4, 24] \times 10^{-4} \text{ (ATLAS 8 TeV, 20.3 fb}^{-1}\text{ [128])}\,,\\
&\hat{Y} \in [-7, 41] \times 10^{-4} \text{ (CMS 8 TeV, 19.7 fb}^{-1}\text{ [129])}\,.
\end{aligned} \tag{3.22}$$

These bounds have been computed assuming SM PDFs. In the rest of this chapter, for this benchmark scenario, we see how the limits based on SM PDFs are modified once a consistent determination of the SMEFT PDFs is performed, requiring a simultaneous fit of the PDFs together with the $\hat{W}$ and $\hat{Y}$ parameters from the high-mass Drell-Yan distributions.

3.4 Data, theory, and fit settings

In Sect. 3.4.1 we present the LHC experimental data that will be used in this chapter for the simultaneous determination of the PDFs and the EFT coefficients from high-mass Drell-Yan cross sections. Subsequently in Sect. 3.4.2, we describe the corresponding theoretical calculations, both in the SM and in the SMEFT benchmark scenario described in Sect. 3.3. In Sect. 3.4.3 we discuss the settings of the baseline SM PDF fit and assess the specific impact of the Run I and Run II high-mass Drell-Yan data on PDFs only. Finally, in Sect. 3.4.4 we outline the fitting methodology adopted for the determination of the PDFs in the SMEFT in this chapter (a different, improved method will be used in Chapter 4), along with their simultaneous determination with the SMEFT Wilson coefficients.

3.4.1 Experimental data

The present analysis is based on the DIS and DY measurements which were part of the strangeness study of [130], which in turn was a variant of the NNPDF3.1 global PDF determination [117], extended with additional high-mass DY cross sections. The DIS structure functions include the same legacy HERA inclusive combination [131] used in the DIS-only joint fit of PDF and SMEFT effects of [101].

No other datasets beyond DIS and DY are considered. In particular, the inclusive jet and top quark production measurements used in [130] are excluded from the present analysis. The rationale behind this choice is the following. As described in Sect. 3.1.2, the SMEFT at dimension-6 level introduces 2499 independent parameters, many of which contribute to the processes used to extract the parton distribution functions. The full PDF fit in the SMEFT (with the consistent power counting in the inverse powers of the new physics scale) is the ultimate future goal of this line of research. Before that, we are forced to make assumptions about the subset of operators and processes involved. The restricted choice of DIS and DY is motivated by the idea that other datasets, such as inclusive jet, could potentially receive corrections from other SMEFT operators, e.g. four-quark operators while being insensitive to the semi-leptonic operators. Including all datasets to effectively determine the PDFs, while considering one or two operators able to impact a subset of processes, would misrepresent the realistic case. We shall investigate the impact of a different collection of SMEFT operators and a different dataset in Chapter 4.

Exp.	$\sqrt{s}$ (TeV)	Ref.	Observable	$n_{\rm dat}$
E886	0.8	[132]	$d\sigma^d_{\rm DY}/d\sigma^p_{\rm DY}$	15
E886	0.8	[133, 134]	$d\sigma^p_{\rm DY}/(dy\, dm_{\ell\ell})$	89
E605	0.04	[135]	$\sigma^p_{\rm DY}/(dx_F\, dm_{\ell\ell})$	85
CDF	1.96	[136]	$d\sigma_Z/dy_Z$	29
D0	1.96	[137]	$d\sigma_Z/dy_Z$	28
D0	1.96	[138]	$d\sigma_{W\to\mu\nu}/d\eta_\mu$ asy.	9
ATLAS	7	[139]	$d\sigma_W/d\eta_l, d\sigma_Z/dy_z$	30
ATLAS	7	[140]	$d\sigma_{Z\to e^+e^-}/dm_{e^+e^-}$	6
ATLAS	7	[141]	$d\sigma_W/d\eta_l, d\sigma_Z/dy_z$	61
ATLAS	7	[142]	$d\sigma_{W+c}/dy_c$	22
ATLAS	8	[143]	$d\sigma_Z/dp_T$	82
ATLAS	8	[144]	$d\sigma_{W+j}/dp_T$	32
CMS	7	[145]	$d\sigma_{W\to l\nu}/d\eta_\ell$ asy.	22
CMS	7	[146]	$d\sigma_{W+c}/dy_c$	5
CMS	7	[146]	$d\sigma_{W^++c}/d\sigma_{W^-+c}$	5
CMS	8	[147]	$d\sigma_Z/dp_T$	28
CMS	8	[148]	$d\sigma_{W\to\mu\nu}/d\eta_\mu$	22
CMS	13	[149]	$d\sigma_{W+c}/dy_c$	5
LHCb	7	[150]	$d\sigma_{Z\to\mu^+\mu^-}/dy_{\mu^+\mu^-}$	9
LHCb	7	[151]	$d\sigma_{W,Z}/d\eta$	29
LHCb	8	[152]	$d\sigma_{Z\to e^+e^-}/dy_{e^+e^-}$	17
LHCb	8	[153]	$d\sigma_{W,Z}/d\eta$	30
Total				**659**

Table 3.1: The low-mass and on-shell Drell-Yan datasets used in the present study. For each dataset we indicate the experiment, the centre-of-mass energy $\sqrt{s}$, the publication reference, the physical observable, and the number of data points

For the purposes of this chapter's study, the DY data can be classified into low-mass, on-shell, and high-mass datasets. Table 3.1 summarises the low-mass and on-shell datasets, where in each case we indicate the experiment, the centre-of-mass energy $\sqrt{s}$, the publication reference, the physical observable, and the number of data points. The only

Exp.	$\sqrt{s}$ (TeV)	Ref.	$\mathcal{L}$ (fb^{-1})	Channel	1D/2D	$n_{\rm dat}$	$m_{\ell\ell}^{\rm max}$ (TeV)
ATLAS	7	[155]	4.9	e^-e^+	1D	13	[1.0, 1.5]
ATLAS (*)	8	[128]	20.3	$\ell^-\ell^+$	2D	46	[0.5, 1.5]
CMS	7	[156]	9.3	$\mu^-\mu^+$	2D	127	[0.2, 1.5]
CMS (*)	8	[129]	19.7	$\ell^-\ell^+$	1D	41	[1.5, 2.0]
CMS (*)	13	[157]	5.1	e^-e^+, $\mu^-\mu^+$ $\ell^-\ell^+$	1D	43, 43 43	[1.5, 3.0]
Total						**270 (313)**	

Table 3.2: Same as Table 3.1 for the neutral-current high-mass Drell-Yan datasets considered in this work. We also indicate the final-state, whether the distribution is 1D (which are differential in the invariant mass, $m_{\ell\ell}$, of the final-state leptons) or 2D (which are differential in both the invariant mass of the leptons, $m_{\ell\ell}$, and in their rapidity, $y_{\ell\ell}$), and the values of $m_{\ell\ell}$ for the most energetic bin. Datasets indicated with **(*)** are used for the first time in this analysis in comparison with [130].

difference as compared to [130] is the removal of the $W \to e\nu$ asymmetry measurements from D0 [154], which were found to be inconsistent with the rest of the Drell-Yan data.

In Table 3.2 we provide the same information as in Table 3.1 but for the neutral-current high-mass Drell-Yan datasets. In Table 3.2 we also indicate the final state, whether the distribution is 1D or 2D (thus differential only in the lepton invariant mass or differential in the lepton invariant mass and rapidity), the integrated luminosity $\mathcal{L}$, and the values of the dilepton invariant mass $m_{\ell\ell}$ for the most energetic bin. We note that while the ATLAS and CMS measurements at $\sqrt{s} = 7$ TeV [155, 156] were already part of the strangeness study of [130], the corresponding 8 TeV and 13 TeV measurements from [128, 129, 157] were not and are considered for the first time in this analysis. For those datasets where data are available in terms of both Born and dressed leptons, the ATLAS 7 TeV analysis being an example thereof, we use the Born data so that it is not necessary to supplement our fixed-order predictions with final-state QED radiation corrections. The CMS 13 TeV data on the other hand are only provided in terms of dressed leptons. In total, there are either 270 or 313 data points in this high-mass category, depending on whether the 13 TeV CMS data are included in the combined channel or in the separate electron and muon channels.

From Table 3.2 one can observe that, with the exception of the CMS 13 TeV data, only one specific leptonic final state is available to be used in the fit. For the CMS 13 TeV measurement instead, one can select between the combined channel or the individual electron and muon final states, which are statistically independent. The separate use of the electron and muon channels is potentially beneficial when considering BSM effects

that are not lepton-flavour universal; however, in the flavour-universal oblique corrections scenario described in Sect. 3.3, it is more convenient to include the data from the combined channel, which displays reduced systematic uncertainties.

3.4.2 Theoretical predictions

We now discuss the settings of the theoretical calculations, both in the SM and in the SMEFT.

SM cross sections. The SM cross sections are computed at next-to-next-to-leading order (NNLO) in QCD and include next-to-leading order (NLO) EW corrections, the latter being especially significant in the high-mass region relevant for this study. In particular, the DIS reduced cross sections (combinations of structure functions) are evaluated at NNLO in the FONLL-C general-mass variable flavour number scheme [158] with `APFEL` [159] interfaced to `APFELgrid` [160]. The Drell-Yan differential distributions are computed using `MCFM` [161] and `amc` [162] interfaced to `APPLgrid` [163] and `APFELgrid` to generate fast NLO interpolation tables which are then supplemented by bin-by-bin K-factors to account for the NNLO QCD and NLO EW corrections. These K-factors are defined as

$$d\sigma_{pp} = \left(d\widehat{\sigma}_{ij}\big|_{\text{NLO QCD}} \otimes \mathcal{L}_{ij}^{\text{NNLO}} \right) \times K_{\text{QCD}} \times K_{\text{EW}} \,, \tag{3.23}$$

where $\otimes$ represents the standard convolution product, $d\sigma_{pp}$ ($d\widehat{\sigma}_{ij}$) is the short-hand notation for the bin-by-bin hadronic cross section (partonic cross section for partons i, j) differential in $m_{\ell\ell}$ (in case of neutral-current (NC) Drell-Yan) or m_T (in case of charged-current (CC) Drell-Yan) and the partonic luminosities $\mathcal{L}_{ij}$ are defined as

$$\mathcal{L}_{ij}(\tau, m) = \int_\tau^1 \frac{dx}{x} \, f_i(x, m^2) f_j(\tau/x, m^2) \,, \tag{3.24}$$

where $m = m_{\ell\ell}$ in the NC case and $m = m_T$ in the CC case and are evaluated at NNLO. The QCD and EW K-factors are defined as

$$K_{\text{QCD}} = \left(\mathcal{L}_{ij}^{\text{NNLO}} \otimes d\widehat{\sigma}_{ij}\big|_{\text{NNLO QCD}} \right) / \left(\mathcal{L}_{ij}^{\text{NNLO}} \otimes d\widehat{\sigma}_{ij}\big|_{\text{NLO QCD}} \right) \,, \tag{3.25}$$

$$K_{\text{EW}} = \left(\mathcal{L}_{ij}^{\text{NNLO}} \otimes d\widehat{\sigma}_{ij}\big|_{\text{NLO QCD+EW}} \right) / \left(\mathcal{L}_{ij}^{\text{NNLO}} \otimes d\widehat{\sigma}_{ij}\big|_{\text{NLO QCD}} \right) \,, \tag{3.26}$$

The NNLO QCD K-factors have been computed using either `MATRIX` [164] or `FEWZ` [165] and cross-checked with the analytic computations of [105, 166]. The NLO EW K-factors have been evaluated with `amc` [162]. Eq. (3.26) accounts also for photon-initiated contributions (using the `NNPDF3.1QED` PDF set [57]) and final-state radiation effects, except when the latter has already been subtracted in the corresponding experimental analysis.

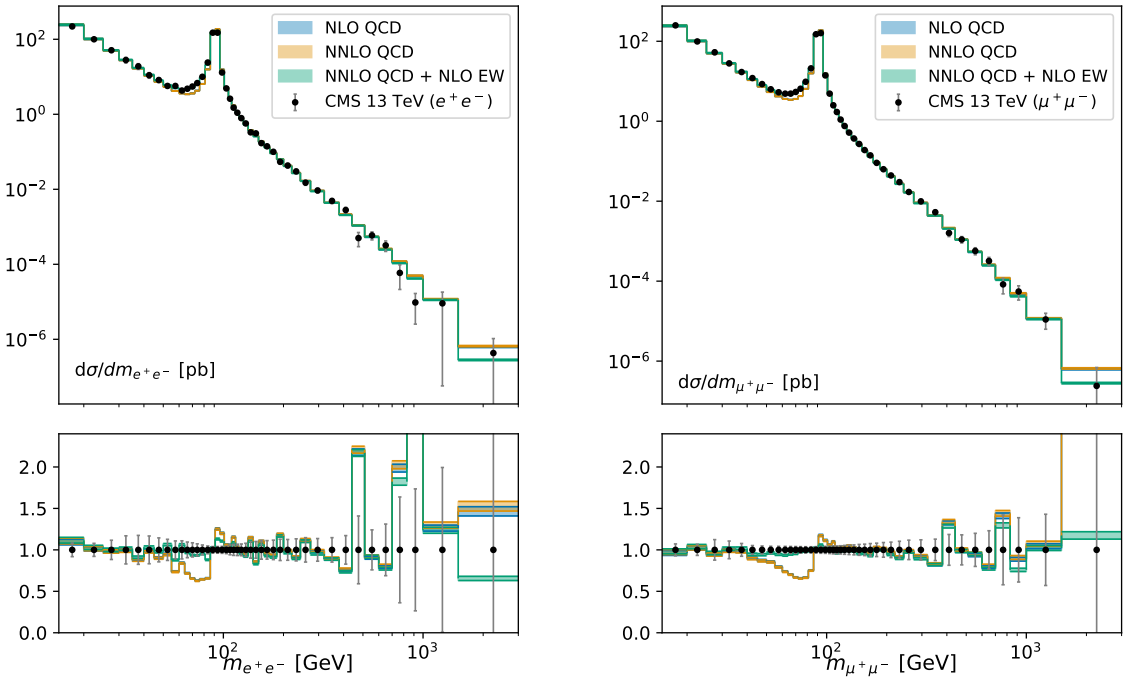


Figure 3.1: Comparison of the CMS Drell-Yan 13 TeV data with the corresponding theoretical calculations at different perturbative orders as a function of the dilepton invariant mass $m_{\ell\ell}$ in the dielectron (left) and dimuon (right panel) final states. The bottom panels display the ratio of the theory calculations to the central value of the experimental data. We display the sum in quadrature of the experimental uncertainties, and the error band in the theory predictions correspond to the one-sigma PDF uncertainties.

Fig. 3.1 displays a comparison between the CMS Drell-Yan distributions at 13 TeV and the corresponding theoretical predictions as a function of the dilepton invariant mass $m_{\ell\ell}$, separately for the dielectron and dimuon final states. The theory calculations are presented at NLO QCD, NNLO QCD, and NNLO QCD combined with NLO EW corrections, in all cases with `NNPDF3.1QED_nnlo_as_0118` as input PDF set, to illustrate the effect of the K-factors of Eq. (3.25) and (3.26). The CMS data are provided in terms of dressed leptons, and hence final state radiation (FSR) QED effects must be included in the electroweak corrections. Accounting for these effects is essential to improve the agreement between theory and data in the region below the Z-mass peak. NLO electroweak corrections are also important in the high-energy tail in $m_{\ell\ell}$, where they are driven by the interplay between (negative) virtual EW effects and (positive) photon-initiated contributions.

A quantitative assessment of the agreement between theoretical predictions and experimental data for the high-mass DY datasets listed in Table 3.2 is presented in Table 3.3, which collects the values of the χ^2 per data point evaluated using the full information on correlated systematics provided by the experimental covariance matrix

$$\chi^2 = \frac{1}{n_{\rm dat}} \sum_{i,j=1}^{n_{\rm dat}} (D_i - T_i) \left(\text{cov}^{-1}\right)_{ij} (D_j - T_j), \tag{3.27}$$

where T_i are the theoretical predictions, D_i the central value of the experimental data and where the multiplicative uncertainties in the experimental covariance matrix (cov_{ij}) are treated using the t_0 prescription as explained in Sect. 1.3.2 and in the references Refs. [35, 167]. One can observe how in general the NNLO QCD corrections are relatively small and that the NLO electroweak ones can be significant, especially for observables presented in terms of dressed leptons (such as the CMS 13 TeV ones) and are required to achieve a good description of the Drell-Yan data in the whole kinematical range available. Note that the input PDF sets used for these calculations include only a subset of these Drell-Yan measurements, in particular only the 7 TeV measurements, for which the data-theory agreement is comparable to the one observed in [117].

The data-theory agreement before including the 8 TeV and 13 TeV data in the PDF fit is generally good, once EW corrections are included, with the exception of the CMS 13 TeV data in the e^+e^- channel, for which the χ^2 per data point remains above 2. As can be observed in Fig. 3.1, the dielectron invariant mass distribution in this channel presents dips at about 500 GeV and 900 GeV which are not present in the $\mu^+\mu^-$ channel. These dips are the origin of this worse data-theory agreement, which is partially reduced once the dataset is included in the fit (see Sect. 3.4.3). We have verified that excluding this dataset from the fit does not change the results of the analysis, and therefore decided to keep it. Further experimental analysis based on the full Runs II and III datasets will tell whether the dips in the distributions in the electron invariant mass will stay.

Dataset	Final state	$n_{\rm dat}$	$\chi^2/n_{\rm dat}$		
			NLO QCD	NNLO QCD	NNLO QCD + NLO EW
ATLAS 7 TeV	e^+e^-	13	1.45	1.77	1.73
ATLAS 8 TeV	$\ell^+\ell^-$	46	1.67	-	1.20
CMS 7 TeV	$\mu^+\mu^-$	127	3.40	1.27	1.54
CMS 8 TeV	$\ell^+\ell^-$	41	2.22	2.21	0.70
CMS 13 TeV	$\ell^+\ell^-$	43	18.7	19.7	1.91
CMS 13 TeV	e^+e^-	43	9.16	9.45	2.32
CMS 13 TeV	$\mu^+\mu^+$	43	15.7	15.8	0.81

Table 3.3: The values of the χ^2 per data point evaluated for the high-mass DY datasets listed in Table 3.2, using theoretical predictions computed at different perturbative accuracy. The PDF sets used here are `NNPDF31_nlo_as_0118`, `NNPDF31_nnlo_as_0118` and `NNPDF31_nnlo_as_0118_luxqed` for the NLO QCD, NNLO QCD and NNLO QCD + NLO EW predictions respectively. For CMS 13 TeV, where different final states are available, we indicate the χ^2 values for each of them. For the ATLAS 8 TeV data, we only evaluated the combined NNLO QCD + NLO EW correction, and hence the pure NNLO QCD result is not given.

SMEFT corrections to the DIS structure functions. In this chapter, we augment the SM calculations of the high-Q^2 DIS reduced cross sections discussed in Ref. [101] and the high-mass Drell-Yan cross sections listed in Table 3.2 with the effects of dimension-six SMEFT operators following the benchmark scenario presented in Sect. 3.3.

Beginning with a discussion of DIS, the SMEFT corrections to the neutral-current deep-inelastic structure functions F_2, F_3 in the benchmark scenario of Sect. 3.3 are obtained by means of a direct calculation in perturbation theory. In order to determine these corrections, we rewrite Eq. (3.20) as the linear combination of four-fermion operators of the form $\bar{q}_\lambda \gamma^\mu q_\lambda \bar{\ell}_{\lambda'} \gamma_\mu \ell_{\lambda'}$, where q_λ is a quark field of helicity λ (with $\lambda = +1$ for a right-handed field and $\lambda = -1$ for a left-handed field) and $\ell_{\lambda'}$ is a lepton field of helicity λ'. The relevant operators for the $\hat{Y}$ parameter are already of this form in Eq. (3.20). For the $\hat{W}$ parameter, the associated operators can be expanded explicitly as:

$$\mathcal{L}_{\rm SMEFT} \supset -\frac{g^2\hat{W}}{4m_W^2}\sum_{i=1}^{3}\Bigg(\bar{e}_L^i\gamma^\mu e_L^i\bar{u}_L^i\gamma_\mu u_L^i - \bar{e}_L^i\gamma^\mu e_L^i\bar{d}_L^i\gamma_\mu d_L^i - \bar{\nu}_L^i\gamma^\mu\nu_L^i\bar{u}_L^i\gamma_\mu u_L^i + \bar{\nu}_L^i\gamma^\mu\nu_L^i\bar{d}_L^i\gamma_\mu d_L^i\Bigg), \tag{3.28}$$

where the index i runs over generations, and the flavour-changing contributions have been

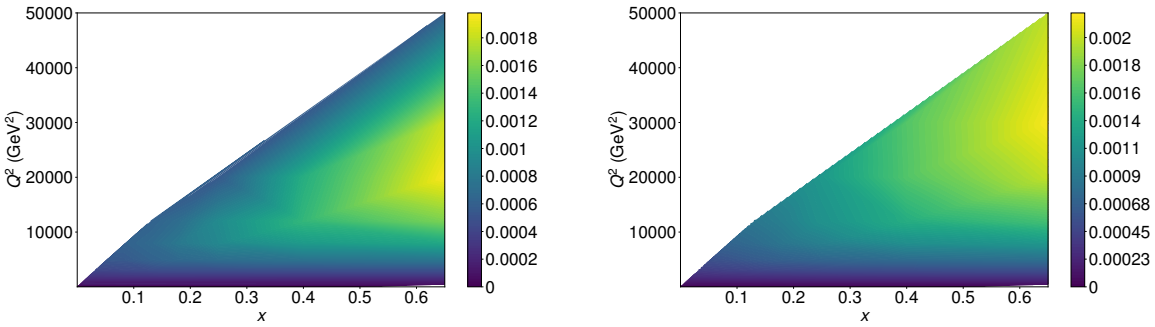


Figure 3.2: Contour maps indicating the value of the EFT correction, $K_{\rm EFT}(\hat{W},\hat{Y})-1$ in Eq. (4.9), for the DIS reduced cross sections as a function of x and Q^2 for two representative values of the EFT parameters: $\hat{W}=-10^{-4}$ (left panel) and $\hat{Y}=-10^{-4}$ (right panel).

dropped since they only contribute to low energy CC structure functions.

The EFT corrections to the DIS structure functions induced by a specific four-fermion operator of the form:

$$\mathcal{L}_{\rm SMEFT} \supset \frac{c^{q\ell}_{\lambda\lambda'}}{\Lambda^2}\bar{q}_\lambda\gamma^\mu q_\lambda\bar{\ell}_{\lambda'}\gamma_\mu\ell_{\lambda'} \tag{3.29}$$

can be shown (by a direct calculation in the parton model, as in Sect. 1.1) to be given by:

$$\Delta F_2(x,Q^2) = \frac{c^{qe}_{\lambda\lambda'}}{\Lambda^2}\frac{Q^2}{2e^2}\left(e_q - K_Z\left(V^e-\lambda' A^e\right)\left(V^q-\lambda A^q\right)\right)\left(xf_q(x,Q^2)+xf_{\bar{q}}(x,Q^2)\right),$$

$$\Delta F_3(x,Q^2) = -\frac{c^{qe}_{\lambda\lambda'}}{\Lambda^2}\frac{Q^2}{2e^2}\left(\lambda\lambda' e_q - K_Z\left(\lambda' V^e-A^e\right)\left(\lambda V^q-A^q\right)\right)\left(f_q(x,Q^2)-f_{\bar{q}}(x,Q^2)\right),$$

where e is the positron charge, e_q is the charge on the quark q in units of the positron charge, θ_W is the Weinberg angle, and $K_Z = Q^2/\sin^2(2\theta_W)(Q^2+m_Z^2)$. The vector and axial couplings are given by $V^e=-\frac{1}{2}+2\sin^2(\theta_W)$, $A^e=-\frac{1}{2}$, $V^q=I_3^q-2\sin^2(\theta_W)e_q$ and $A^q=I_3^q$, where I_3^q is the third component of the quarks' weak isospin. These formulae are the natural generalisations of those derived in [101], where only right-handed four-fermion operators were considered. Taking combinations of these DIS structure-function corrections according to Eq. (3.20) for the $\hat{Y}$ parameter and to Eq. (3.28) for the $\hat{W}$ parameter yields the sought-for EFT corrections for DIS observables.

This calculation has been implemented in `APFEL` [159] following the strategy presented in [101]. FK tables are produced from `APFEL` [160] and then used to evaluate the DIS K-factors defined in Eq. (3.23). Furthermore, we have used `APFEL` to include the higher-order QCD corrections in the SMEFT sector, so that in fact Eq. (3.23) holds exactly for the DIS K-factors in our study.

Fig. 3.2 displays contour maps indicating the EFT correction, $K_{\rm EFT}(\hat{W},\hat{Y})-1$ in Eq. (4.9), for the DIS reduced cross sections (which include both F_2 and xF_3) as a function of x and Q^2 for two representative values of the EFT parameters, $\hat{W}=-10^{-4}$ and

$\hat{Y} = -10^{-4}$. These maps should be compared with Fig. 1 of [101], which considered different EFT scenarios. We find that the overall effect of non-zero $\hat{W}$ and $\hat{Y}$ parameters is rather small, well below the percent level even for the highest bins in Q^2 covered by the HERA data. This comparison highlights how, in this benchmark EFT scenario, the constraints on the $\hat{W}$ and $\hat{Y}$ parameters will be completely dominated by the (high-mass) Drell-Yan cross sections.

SMEFT corrections for Drell-Yan distributions. Similarly to DIS, the SMEFT corrections are negligible for dilepton invariant masses of $m_{\ell\ell} \leq 200$ GeV and hence we can safely adopt the SM calculations there. For the high-mass Drell-Yan distributions however, we must include the SMEFT corrections, which are appreciable.

In a similar manner as for higher-order QCD and EW corrections, we can define correction factors that encapsulate the linear and quadratic modifications induced by the dimension-six SMEFT operators. Adopting an operator normalisation such that:

$$\mathcal{L}_{\rm SMEFT} = \mathcal{L}_{\rm SM} + \sum_{n=1}^{n_{\rm op}} \frac{c_n}{v^2} \mathcal{O}_n \,, \tag{3.30}$$

with $n_{\rm op}$ indicating the number of operators that contribute to a given benchmark scenario and c_n being the (dimensionless) Wilson coefficient associated to $\mathcal{O}_n$, the linear EFT corrections can be parametrised as:

$$R^{(n)}_{\rm SMEFT} \equiv \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}^{(n)}_{ij,\rm SMEFT}\right) / \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}_{ij,\rm SM}\right) \,, \quad n = 1 \ldots, n_{\rm op} \,, \tag{3.31}$$

with $\mathcal{L}^{\rm NNLO}_{ij}$ being the usual partonic luminosity evaluated at NNLO QCD, $d\widehat{\sigma}_{ij,\rm SM}$ the bin-by-bin partonic SM cross section, and $d\widehat{\sigma}^{(n)}_{ij,\rm SMEFT}$ the corresponding partonic cross section associated to the interference between $\mathcal{O}_n$ and the SM amplitude $\mathcal{A}_{\rm SM}$ when setting $c_n = 1$. Likewise, the ratio encapsulating the quadratic effects is defined as:

$$R^{(n,m)}_{\rm SMEFT} \equiv \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}^{(n,m)}_{ij,\rm SMEFT}\right) / \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}_{ij,\rm SM}\right) \,, \quad n, m = 1 \ldots, n_{\rm op} \,, \tag{3.32}$$

with the bin-by-bin partonic cross section $d\widehat{\sigma}^{(n,m)}_{ij,\rm SMEFT}$ now being evaluated from the squared amplitude $\mathcal{A}_n\mathcal{A}_m$ associated to the operators $\mathcal{O}_n$ and $\mathcal{O}_m$ when $c_n = c_m = 1$. The partonic cross sections in these ratios are computed at LO. In terms of Eqns. (4.5) and (4.7), we can define the EFT K-factors as:

$$K_{\rm EFT} = 1 + \sum_{n=1}^{n_{\rm op}} c_n R^{(n)}_{\rm SMEFT} + \sum_{n,m=1}^{n_{\rm op}} c_n c_m R^{(n,m)}_{\rm SMEFT} \,, \tag{3.33}$$

which allow us to express general Drell-Yan or DIS cross sections accounting for the

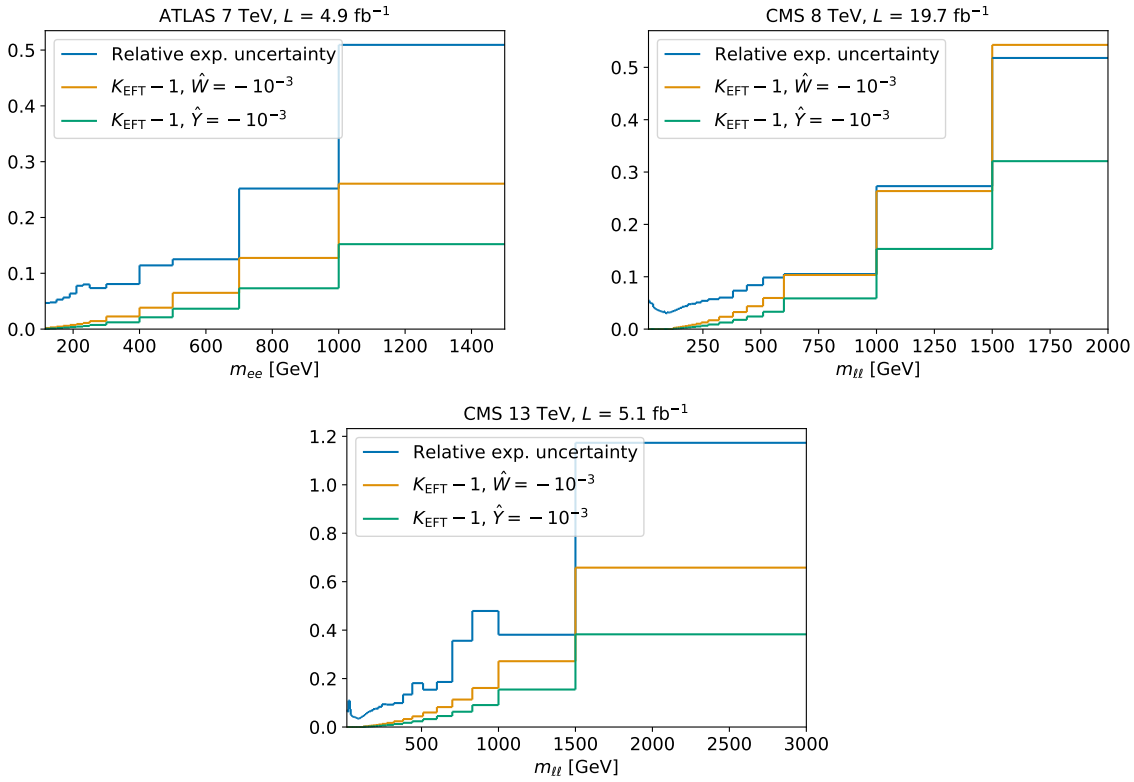


Figure 3.3: Comparison between the (relative) experimental uncertainties and the corresponding EFT corrections, $K_{\rm EFT}(\hat{W},\hat{Y})-1$ in Eq. (4.9), for the ATLAS 7 TeV, CMS 8 TeV, and CMS 13 TeV Drell-Yan $m_{\ell\ell}$ distributions, for two representative values of $\hat{W}$ and $\hat{Y}$.

dimension-six operators in Eq. (4.4) as:

$$d\sigma_{\rm SMEFT} = d\sigma_{\rm SM} \times K_{\rm EFT} \tag{3.34}$$

where the $d\sigma_{\rm SM}$ is the state-of-the-art SM prediction including NNLO QCD and NLO EW corrections. In this approach, the SMEFT predictions inherit factorisable higher-order radiative correction [126, 127]. The SMEFT K-factors in Eq. (4.9) are precomputed before the fit using a reference SM PDF set and then kept fixed. The effect of varying the input NNLO PDF in Eqns. (4.5) and (4.7) is quantitatively assessed in App. C of Ref. [102] and it is found to be at the permil level. As a result, this effect will be neglected in the following.

Fig. 3.3 illustrates the size of the EFT corrections in the benchmark scenario from Sect. 3.3 by comparing $(K_{\rm EFT}-1)$ with the relative experimental uncertainties for the ATLAS 7 TeV, CMS 8 TeV, and CMS 13 TeV Drell-Yan $m_{\ell\ell}$ distributions. We provide results for two representative points in the $(\hat{W},\,\hat{Y})$ parameter space, namely $(\hat{W},\,\hat{Y}) = (10^{-3},0)$ and $(0,10^{-3})$. One can observe how for these values of $(\hat{W},\,\hat{Y})$, and particularly for the ATLAS 8 TeV data, the SMEFT corrections to the Drell-Yan cross sections become comparable with the experimental uncertainties, increasing steadily with $m_{\ell\ell}$.

3.4.3 Baseline SM PDFs

The settings for this baseline SM PDF fit used in this chapter are the same as those used in the strangeness study of [130], itself a variant of NNPDF3.1 [117]. As described in Sect. 3.4.1, in this work we consider only DIS and Drell-Yan datasets, with the latter augmented as compared to [130] with the new high-mass measurements indicated in Table 3.2.

In general, the fit quality of the baseline SM PDF set is similar to that of the global fit of [130], although the description of the CMS 13 TeV invariant mass distribution in the combined electron and muon channels remains sub-optimal. See App. B of Ref. [102] for a complete discussion of the fit quality.

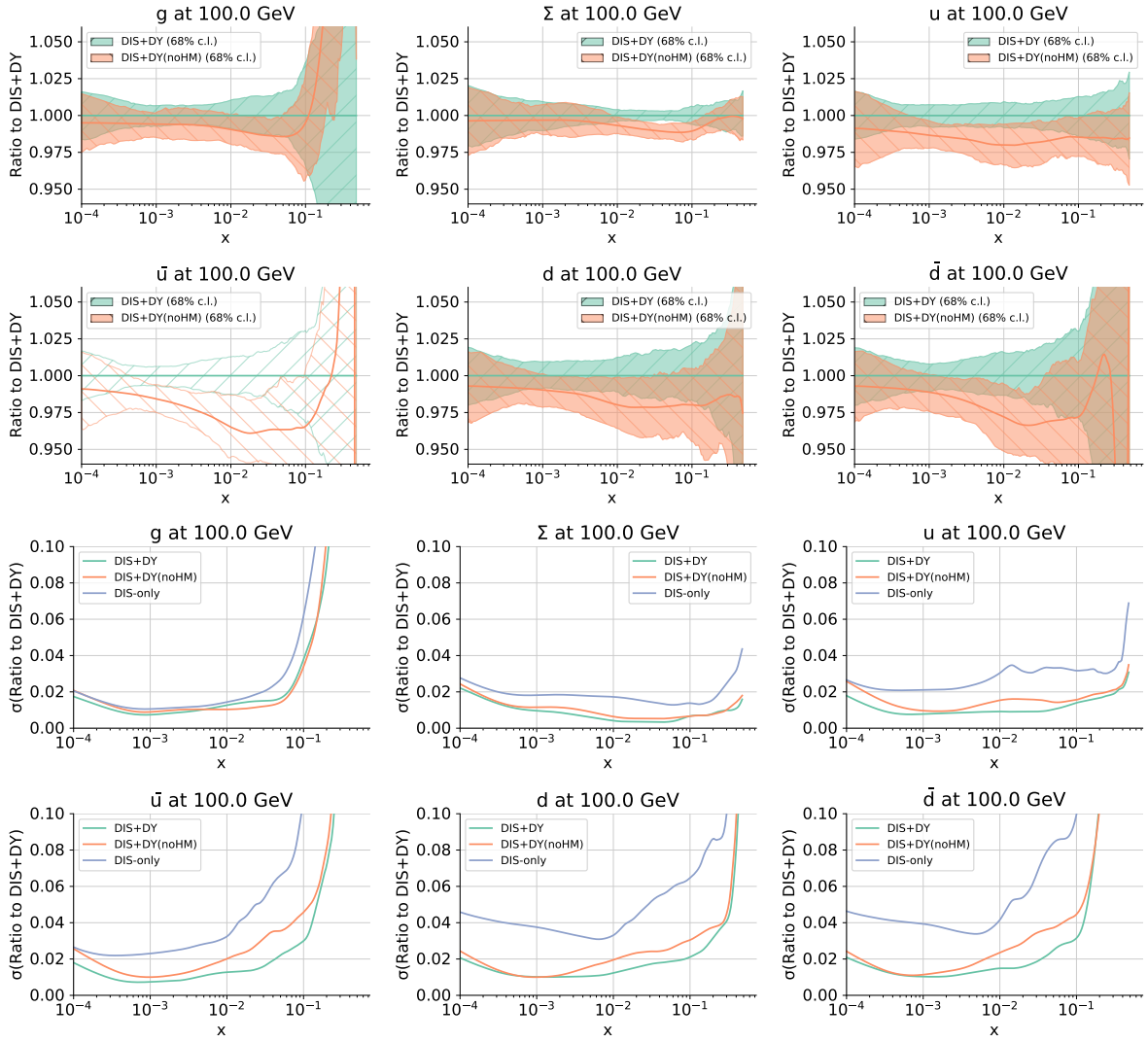


Figure 3.4: Comparison between the baseline SM PDF set of this work, labelled 'DIS+DY', with the corresponding fit without high-mass DY data. We show results at $Q = 100$ GeV for PDFs normalised to the central value of the baseline (upper) and for the relative PDF uncertainties (lower panels). In the latter case, we also display the PDF uncertainties from the DIS-only fit.

Fig. 3.4 displays a comparison between this baseline SM PDF set, labelled 'DIS+DY', with the same fit but without any datapoints from the high-mass DY datasets listed in Table 3.2, labelled 'DIS+DY (no HM)'. We show results at $Q = 100$ GeV both for the PDFs normalised to the central value of the baseline and for the relative PDF uncertainties. In the latter case we also display the PDF uncertainties from a corresponding DIS-only fit. The latter comparison shows that the DY cross sections significantly reduce the PDF uncertainties of the DIS-only fit. The addition of the high-mass DY data leads to a visible uncertainty reduction in the $0.005 \lesssim x \lesssim 0.3$ region as compared to the 'DIS+DY(noHM)' reference as well an upwards shift of the up and down quarks and antiquark PDF.

We therefore find that the available high-mass DY data can have an appreciable impact on the light quark and antiquark PDFs, despite the fact that in terms of Run II data our analysis is restricted to a single low-luminosity high-mass DY dataset. Yet more stringent constraints on the PDFs are expected from the measurements based on the full Run II and Run III datasets, as well as from those to be provided by the HL-LHC [119]. We study the anticipated impact of the HL-LHC measurements in Sect. 3.6.

3.4.4 Methodology for the simultaneous PDF and EFT fits

Let us denote by $\mathbf{c} = (c_1, c_2, \ldots, c_{N_{\rm op}})$ the array containing the Wilson coefficients associated to the $N_{\rm op}$ dimension-six operators contributing to a given SMEFT scenario, where c_n are defined as in Eq. (4.4). For each point $\mathbf{c}_i$ in the scan of the EFT parameter space, we evaluate the Drell-Yan and the DIS cross sections as described in Sect. 4.2. Subsequently, we determine the best-fit PDFs associated to $\mathbf{c}_i$ by means of the standard NNPDF methodology, which determines the minimum of the χ^2 in the space of the PDF parameters (subject to cross-validation, to avoid overlearning). We note that this χ^2, defined in Eq. (3.27), keeps fully into account the experimental systematic correlations among all the measurements D_i included in the PDF analysis.

This procedure results in a sampling of the χ^2 values in the EFT parameter space, which we denote by $\chi^2_{\rm eftp}(\mathbf{c}_i)$ (as in: EFT-PDFs). Alternatively, one could also evaluate the same DIS and DY cross sections using instead the baseline SM PDF set, ending up with χ^2 values which we denote by $\chi^2_{\rm smp}(\mathbf{c}_i)$ (as in: SM-PDFs). The comparison between the resulting bounds on the EFT coefficients obtained from $\chi^2_{\rm eftp}(\mathbf{c}_i)$ and from $\chi^2_{\rm smp}(\mathbf{c}_i)$ quantifies the relevance of producing consistent joint determinations of PDFs and Wilson coefficients when studying EFTs in high-energy tails. This strategy follows the one adopted in the proof-of-concept DIS-only study [101], now extended to LHC processes.

Close enough to a local minimum $\chi^2_0 = \chi^2\left(\mathbf{c}^{(0)}\right)$ associated with best-fit values $\mathbf{c}^{(0)}$,

the χ^2 as a function of the EFT coefficients can be approximated by a quadratic form

$$\chi_i^2 \equiv \chi^2(\mathbf{c}_i) = \chi_0^2 + \sum_{n,m=1}^{N_{\rm op}} \left(c_{n,i} - c_n^{(0)}\right) H_{nm} \left(c_{m,i} - c_m^{(0)}\right) , \tag{3.35}$$

with H_{nm} being the usual Hessian matrix in the EFT parameter space. Restricting the EFT calculations to their linear, $\mathcal{O}\left(\Lambda^{-2}\right)$, contributions, Eq. (3.35) becomes exact in the case of $\chi^2_{\rm smp}(\mathbf{c}_i)$ (where cross sections are evaluated with SM PDFs). The reason is that in this case all dependence on the EFT coefficients is encoded in the partonic cross sections.

However, this is not true for $\chi^2_{\rm eftp}(\mathbf{c}_i)$, since now there will be a (non-linear) EFT back-reaction onto the PDFs and hence Eq. (3.35) is only valid up to higher orders in the EFT expansion, even if the EFT cross sections themselves are evaluated in the linear approximation. Eq. (3.35) can thus be only considered a reasonable approximation in the case that the SMEFT PDFs are not too different from their SM counterparts.

Hence, if we work with linear EFT calculations, provided the sampling in the EFT parameter space is sufficiently broad and fine-grained, and that the EFT-induced distortion on the PDFs is moderate, we can extract the parameters χ_0^2 and $\mathbf{c}^{(0)}$ and the Hessian matrix H using least-squares regression from Eq. (3.35), using $\chi^2_{\rm smp}$ for the SM PDFs and $\chi^2_{\rm eftp}$ for the SMEFT PDFs. The associated confidence level contours are determined by imposing

$$\Delta\chi^2(\mathbf{c}) \equiv \chi_i^2(\mathbf{c}) - \chi_0^2 = \sum_{n,m=1}^{N_{\rm op}} \left(c_n - c_n^{(0)}\right) H_{nm} \left(c_m - c_m^{(0)}\right) = \text{constant} , \tag{3.36}$$

where this constant depends on the number of degrees of freedom. For linear EFT two-parameter fits, such as those in the benchmark scenario, in the context of the HL-LHC projections we shall eventually consider, imposing Eq. (3.36) leads to elliptic contours in the $(\hat{W}, \hat{Y})$ plane.

To conclude this section, we give details on how we account for PDF uncertainties and the statistical uncertainty associated to the finite replica sample of the NNPDF Monte Carlo sets that we use here.

PDF uncertainty. In Sects. 3.5 and 3.6.3 we will present bounds on the EFT parameters using the SM PDFs with and without the PDF uncertainties being accounted for. In order to estimate these, we follow the procedure detailed above to determine the confidence level intervals for the EFT parameters but now using the kth Monte Carlo replica of the PDF set, rather than the central replica $k = 0$ as done when PDF uncertainties are neglected.

One ends up with $N_{\rm rep}$ values of the upper and lower bounds:

$$\left[\mathbf{c}_{\rm min}^{(k)}, \mathbf{c}_{\rm max}^{(k)}\right] , \qquad k = 1, \ldots, N_{\rm rep} , \tag{3.37}$$

and then the outermost bounds in the 68% envelope are considered to be the bounds on the EFT parameters $\mathbf{c}$, now including the 1σ-PDF uncertainty. This is very important to account for, given that in the case of the bounds determined using $\chi^2_{\rm eftp}$, the PDF uncertainty is already included by construction, given that the Wilson coefficients are determined from the global set of PDFs, exactly as in the case of the α_s determination from a global set of PDFs of [168, 169]. A more sophisticated way to extract parameters such as α_s of the Wilson coefficients from a global fit of PDFs, that includes the correlations between these parameters and the PDFs, is given by the correlated replica method proposed in the more recent α_s determination in [42]. The latter would allow better accounting of the correlations between Wilson coefficients and PDFs. However we do not use it here due to the fact that the correlations of the PDFs with the Wilson coefficients are much smaller than those with the strong coupling constant, and due to its large computational cost. This issue is addressed with the introduction of a new methodology, presented in Chapter 4.

Methodological uncertainty. In a simultaneous fit of PDFs and EFT coefficients, for each set of Wilson coefficients $\mathbf{c}_i$ one has a PDF fit composed of $N_{\rm rep}$ Monte Carlo replicas. The major methodological uncertainty is associated to finite-$N_{\rm rep}$ effects, and can be estimated by bootstrapping across the replicas, as explained in the $\alpha_s(m_Z)$ extraction of [42]. Specifically, for each value of $\mathbf{c}_i$ we perform $N_{\rm res}$ re-samples of all $N_{\rm rep}$ replicas with replacement, and compute the theory predictions:

$$\mathbf{T}^{(\rm res)}_{i,lk} , \quad \begin{array}{l} l = 1, \ldots, N_{\rm res} \\ k = 1, \ldots, N_{\rm rep} \end{array} , \tag{3.38}$$

such that there are $N_{\rm res}$ re-samples each composed of an $N_{\rm rep}$-sized array of theory predictions. Since this re-sampling is done with replacement, it differs from the original sample in that it contains duplicates and missing values. The average theory prediction is then obtained for each of these bootstrapped sets:

$$\overline{\mathbf{T}}_{i,l} = \left\langle \mathbf{T}^{(\rm res)}_{i,lk} \right\rangle_{\rm rep} , \quad l = 1, \ldots, N_{\rm res} . \tag{3.39}$$

These bootstrapped theory predictions $\overline{\mathbf{T}}_{i,l}$ are used to evaluate the χ^2 to data, with the finite-size uncertainty given by the standard deviation across each bootstrap re-sample:

$$\sigma_{\chi^2_i} = {\rm std}\left(\chi^2_{i,l}\right)\Big|_{\rm res} . \tag{3.40}$$

A value of $N_{\rm res} \simeq 10^4$ re-samples is found to be sufficient to achieve stable results for the estimate of the finite-size uncertainties defined by Eq. (3.40).

3.5 Results from current Drell-Yan data

In this section, we present results for the SMEFT PDFs extracted from DIS and Drell-Yan data in the benchmark SMEFT scenario. We compare them with their SM counterparts at the level of partonic luminosities and assess how the bounds obtained on the $\hat{W}$ and $\hat{Y}$ parameters in this simultaneous SMEFT and PDF fit compare to those based on assuming SM PDFs. We present results for one-dimensional fits where only one of the $\hat{W}$ or the $\hat{Y}$ parameter is allowed to be non-zero; the reason for this choice is that, in a fit including only high-mass neutral-current Drell-Yan processes, there exists a flat direction when $\hat{W}$ and $\hat{Y}$ are varied simultaneously, since both operators scale as $O(q^4)$ and thus cannot both be constrained by a single 1D distribution. This degeneracy can only be lifted once high-mass charged-current DY data is included in the fit. As we demonstrate in Sect. 3.6, thanks to the HL-LHC it will be possible to carry out a simultaneous fit of the PDFs and the two EFT parameters $(\hat{W}, \hat{Y})$.

Taking into account the existing bounds reported in Sect. 3.3, as well as the sensitivity of available high-mass Drell-Yan data to the EFT coefficients illustrated by Fig. 3.3, here we have adopted the following sampling ranges for the $\hat{W}$ and $\hat{Y}$ parameters:

$$\hat{W} \times 10^4 \in [-22, 14]\,, \qquad \hat{Y} \times 10^4 \in [-20, 20]\,. \tag{3.41}$$

We used 21 sampling values of $\hat{Y}$ equally spaced in this interval, hence in steps of $\Delta\hat{Y} = 2 \times 10^{-4}$. In the case of $\hat{W}$ it was found convenient to instead use 15 points equally spaced between -14×10^{-4} and 14×10^{-4} in steps of $\Delta\hat{W} = 2 \times 10^{-4}$, and then to add two more values at $\hat{W} = -18 \times 10^{-4}$ and -22×10^{-4}.

Fig. 3.5 displays the obtained values of $\Delta\chi^2$, Eq. (3.36), as a function of $\hat{W}$ and $\hat{Y}$ in the case of the SMEFT PDFs (that is, using the values of $\chi^2_{\rm eftp}(\mathbf{c}_i)$). These χ^2 values are evaluated as a sum over those datasets from Table 3.1 and 3.2 that receive non-zero EFT corrections, namely the DIS datasets that have a reach in Q^2 above $(120)^2$ GeV^2 (namely HERA and NMC), and the ATLAS and CMS high-mass Drell-Yan measurements in Table 3.2. Furthermore, only linear EFT effects are included in the calculation of the DIS and DY cross sections, while the (subleading) quadratic corrections are neglected in this scenario. The error bars in the $\Delta\chi^2_i$ points of Fig. 3.5 indicate the methodological finite-size uncertainties evaluated with the bootstrapping method described in Sect. 3.4 and the horizontal line corresponds to the $\Delta\chi^2 = 4$ condition associated to a 95% CL

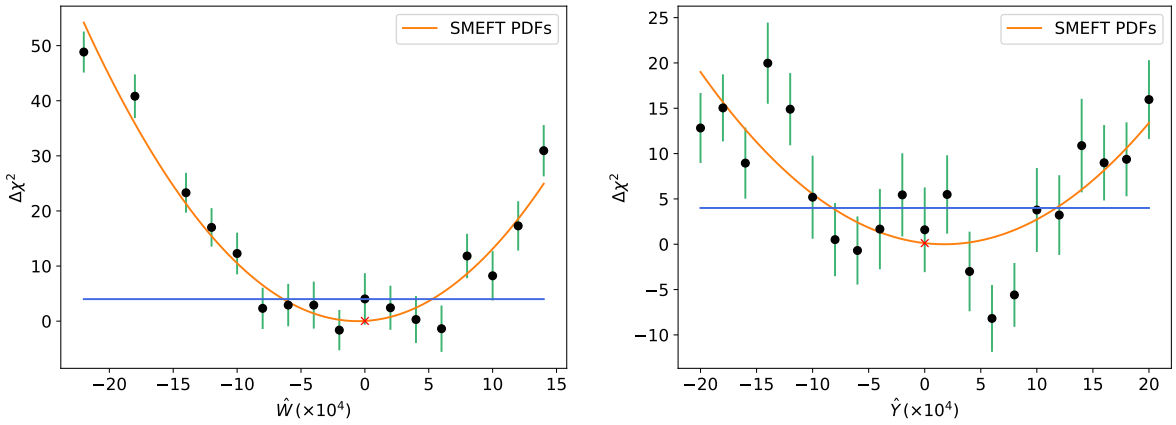


Figure 3.5: The values of $\Delta\chi^2$, Eq. (3.36), obtained for the SMEFT PDFs (thus using the $\chi^2_{\rm eftp}(\mathbf{c}_i)$ values) as a function of $\hat{W}$ (left) and $\hat{Y}$ (right panel) in the sampling ranges of Eq. (3.41) together with the corresponding parabolic fits. The error bars indicate the finite-size uncertainties and the horizontal line corresponds to the $\Delta\chi^2 = 4$ condition defining the 95% CL intervals. The red cross indicates the SM expectation, $\hat{W} = \hat{Y} = 0$.

interval. We also show in Fig. 3.5 the results of the associated parabolic fits,

$$\Delta\chi^2(\hat{W}) = \left(\hat{W} - \hat{W}^{(0)}\right)^2 / \left(\delta\hat{W}\right)^2 , \tag{3.42}$$

and likewise for $\Delta\chi^2(\hat{Y})$. From the results in Fig. 3.5, one observes that both the $\hat{W}$ and $\hat{Y}$ parameters agree with the SM expectation within uncertainties.

Fig. 3.6 then compares the results of the parabolic fits based on the SMEFT PDFs as displayed in Fig. 3.5 with their counterparts obtained in the case of the SM PDFs. That is, in the latter case one carries out parabolic fits to the $\chi^2_{\rm smp}$ values, as is customary in the literature for the EFT analyses. The insets highlight the region close to $\Delta\chi^2 \simeq 0$. For the $\hat{W}$ parameter, the consistent use of SMEFT PDFs leaves the best-fit value essentially unchanged but increases the coefficient uncertainty $\delta\hat{W}$, leading to a broader parabola. Similar observations can be derived for the $\hat{Y}$ parameter, though here one also finds a upwards shift in the best-fit values by $\Delta\hat{Y} \simeq 2\times 10^{-4}$ in addition to a parabola broadening, when SMEFT PDFs are consistently used. We note that the SM PDF parabolas in Fig. 3.6 are evaluated using the central PDF replica and hence do not account for PDF uncertainties.

Table 3.4 summarises the 68% and 95% CL bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the corresponding parabolic $\Delta\chi^2$ fits using either the SM or the SMEFT PDFs shown in Fig. 3.6. The fourth and fifth column indicate the absolute shift in best-fit values and the percentage broadening of the fit parameter uncertainties when the SMEFT PDFs are consistently used instead of the SM PDFs (either without or with PDF

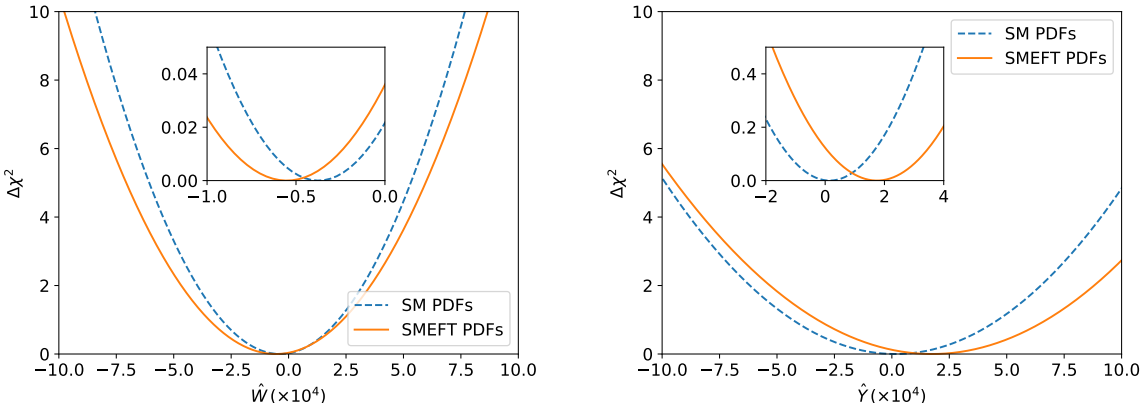


Figure 3.6: Comparison between the results of the parabolic fits to $\Delta\chi^2$, Eq. (3.42), for the $\hat{W}$ (left) and $\hat{Y}$ (right panel) parameters for either the SMEFT PDFs ($\chi^2_{\rm eftp}$, already displayed in Fig. 3.5) or the SM PDFs (hence with $\chi^2_{\rm smp}$). The insets zoom on the region close to $\Delta\chi^2 \simeq 0$.

uncertainties):

$$\text{best fit shift} \equiv \left(\hat{W}^{(0)}\Big|_{\rm SMEFT\,PDF} - \hat{W}^{(0)}\Big|_{\rm SM\,PDF}\right) , \tag{3.43}$$

$$\text{broadening} \equiv \left(\delta\hat{W}^{(0)}\Big|_{\rm SMEFT\,PDF} - \delta\hat{W}^{(0)}\Big|_{\rm SM\,PDF}\right) \Big/ \delta\hat{W}^{(0)}\Big|_{\rm SM\,PDF} , \tag{3.44}$$

and likewise for the $\hat{Y}$ parameter.

In the specific case of the SM PDF results, Table 3.4 indicates the bounds obtained without (upper) and with (lower entry) PDF uncertainties accounted for; recall that the SMEFT PDF bounds already include PDF uncertainties by construction (see Sect. 3.4.4). By comparing the bounds obtained when PDF uncertainties are accounted for to those neglecting PDF uncertainty, one observes a systematic broadening of the bounds from both the lower and upper limits, as was also reported in [101].

When PDF uncertainties are neglected (accounted for) when using the SM PDFs to constrain the EFT parameters, the consistent use of the SMEFT PDFs leads to both a shift in the best-fit values of magnitude $\Delta\hat{W} = -2 \times 10^{-5}$ and $\Delta\hat{Y} = +1.6 \times 10^{-4}$ as well as to an increase (decrease) of the fit parameter uncertainties, with $\delta\hat{W}$ and $\delta\hat{Y}$ growing by 15% and 12% (decreasing by 11% and 13%) respectively. This result shows that, given available Drell-Yan data and once PDF uncertainties are accounted for, the bounds on the EFT parameters are actually *improved* once SMEFT PDFs are adopted.

All in all, the effect of the consistent treatment of the SMEFT PDFs in the interpretation of high-mass DY cross sections is moderate but not negligible, either loosening or tightening up the obtained bounds on the EFT parameters (depending on whether or not PDF uncertainties are accounted for to begin with) by up to 15% and, in the case of $\hat{Y}$ parameter, shifting its central value by one-third of the 68% CL parameter uncertainty.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$\hat{W} \times 10^4$ (68% CL)	$[-3.0, 2.2]$ $[-4.3, 3.8]$	$[-3.5, 2.4]$	-0.2 -0.3	$+13\%$ -27%
$\hat{W} \times 10^4$ (95% CL)	$[-5.5, 4.7]$ $[-6.8, 6.3]$	$[-6.4, 5.3]$	-0.2 -0.3	$+15\%$ -11%
$\hat{Y} \times 10^4$ (68% CL)	$[-4.4, 4.7]$ $[-6.7, 7.5]$	$[-3.4, 6.9]$	$+1.6$ $+1.4$	$+13\%$ -27%
$\hat{Y} \times 10^4$ (95% CL)	$[-8.8, 9.2]$ $[-11.1, 12.0]$	$[-8.3, 11.8]$	$+1.6$ $+1.3$	$+12\%$ -13%

Table 3.4: The 68% CL and 95% CL bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the corresponding parabolic fits to the $\Delta\chi^2$ values calculated from either the SM or the the SMEFT PDFs. For the SM PDF results, we indicate the bounds obtained without (upper) and with (lower entry) PDF uncertainties accounted for; the SMEFT PDF bounds already include PDF uncertainties by construction, while the methodological uncertainty is included according to the approached described in Sect. 3.4. The fourth and fifth column indicate the absolute shift in best-fit values, Eq. (3.43) and the percentage broadening of the EFT parameter uncertainties, Eq. (3.44), when the SMEFT PDFs are consistently used instead of the SM PDFs.

Such a relatively moderate effect can be partly understood from the limited availability of high-mass DY measurements for EFT interpretations, with a single dataset at 13 TeV, and even in this case, with it being restricted to a small fraction of the Run II luminosity. As we will demonstrate in Sect. 3.6, the impact of SMEFT PDFs becomes much more significant once higher-statistics measurements of the NC and CC Drell-Yan tails become available at the HL-LHC, loosening the bounds on $\hat{W}$ and $\hat{Y}$ by up to a factor 5.

We now move to assess how the SMEFT PDFs relate to their SM counterparts, and determine the extent to which it is possible to reabsorb EFT effects into the PDFs. Fig. 3.7 displays a comparison between the SM and the SMEFT PDF luminosities for representative values of the $\hat{W}$ (upper) and $\hat{Y}$ (lower panel) parameters. The values of $\hat{W}$ and $\hat{Y}$ are chosen to be close to the upper and lower limits of the 95% CL intervals reported in Table 3.4. The error band in the SM PDFs corresponds to the 68% CL PDF uncertainty, while for the SMEFT PDFs only the central values are shown.

In all cases, one finds that the EFT-induced shifts on the luminosities are smaller than their standard deviation. The biggest differences, relative to uncertainties, are observed in the quark-antiquark luminosities for $m_X \gtrsim 500$ GeV. This finding can be understood from the fact that the NC Drell-Yan cross section is proportional to the $u\bar{u}$ and $d\bar{d}$ combinations at leading order, but the up and down quark PDFs are already well constrained by lower-energy DIS measurements. Furthermore, we have verified that the size of the PDF

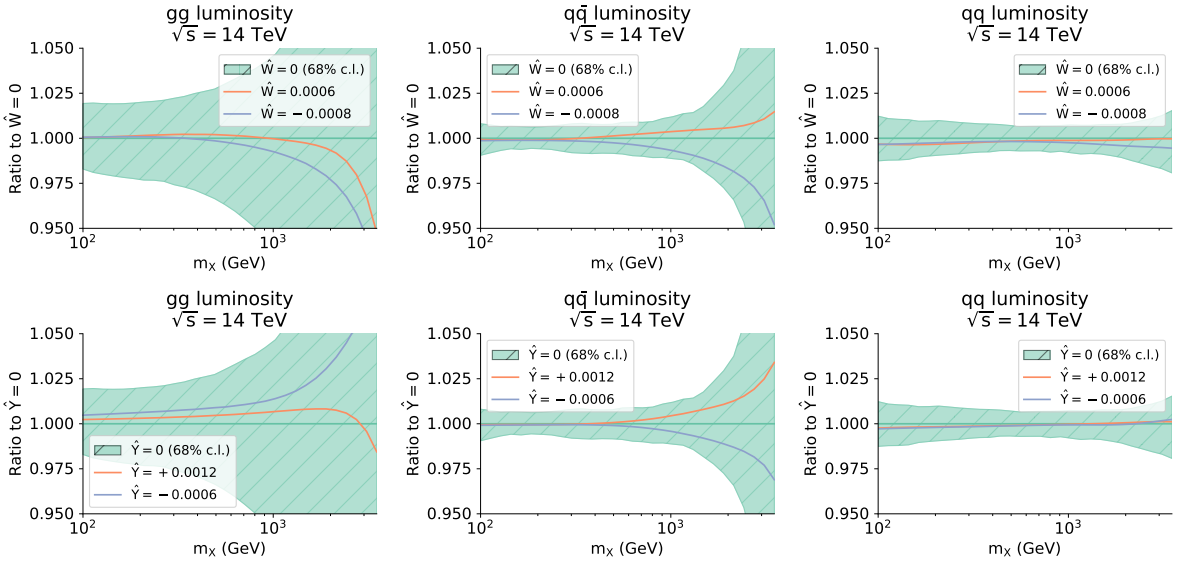


Figure 3.7: Comparison between the SM PDF luminosities with their SMEFT counterparts, displayed as ratios to the central value of the SM luminosities, for representative values of the $\hat{W}$ (upper) and $\hat{Y}$ (lower panel) parameters. The values of $\hat{W}$ and $\hat{Y}$ are chosen to be close to the upper and lower limits of the 95% CL intervals reported in Table 3.4.

uncertainties is unchanged in the SMEFT fits. The results of Fig. 3.7 are consistent with those of Table 3.4 and demonstrate that, with current data, the interplay between EFT effects and PDFs in the high-mass Drell-Yan tails is appreciable but remains subdominant as compared to other sources of uncertainty.

One important question in this context concerns how one could disentangle the EFT-induced shifts in the PDF luminosities displayed in Fig. 3.7 from other possible sources of deviations, such as internal inconsistencies in some datasets or missing higher orders in the SM calculations. A dedicated study to this end is performed in Chapter 5.

3.6 Results from projected HL-LHC Drell-Yan data

The results presented in the previous section indicate that, given the available unfolded Drell-Yan measurements, the impact of a simultaneous determination of the PDFs together with the EFT parameters remains moderate. However, it is conceivable that this interplay between PDFs and BSM effects in the high-energy tails of Drell-Yan cross sections will become more significant once more data are accumulated. With this motivation, we revisit the analysis of Sect. 3.5 now accounting for the impact of projected High-Luminosity LHC pseudo-data generated for the present study. We demonstrate that a consistent joint determination of PDFs is crucial for EFT studies at the HL-LHC.

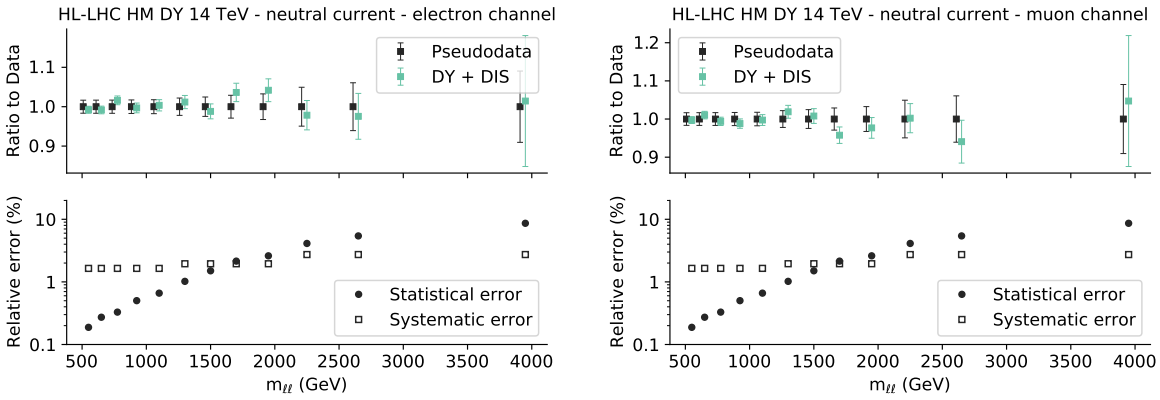


Figure 3.8: Top panels: comparison of the projected HL-LHC pseudo-data for high-mass neutral-current Drell-Yan in the dielectron (left) and dimuon (right) final states as a function of $m_{\ell\ell}$ with the corresponding theory predictions obtained from the SM PDF baseline. The theoretical predictions, generated according to Eq. (3.45), are accompanied by their corresponding PDF uncertainties (green bars). Lower panels: the percentage statistical and systematic uncertainty in each $m_{\ell\ell}$ bin of the HL-LHC pseudo-data.

3.6.1 Generation of HL-LHC pseudo-data

The HL-LHC pseudodata in this chapter is generated in the same way as in Sect. 2.3.3 of Chapter 2. However, here we also generate data for charged-current DY data, whereas in Sect. 2.3.3 we used only neutral-current DY data. Additionally, the PDF set used as an input to generate the theoretical prediction on which the pseudodata is based is the DIS+DY baseline that was presented in Sect. 3.4.3 (rather than the `luxQED` set used in Sect. 2.3.3).

In the case of the CC pseudo-data, the lack of unfolded measurements of the m_T distribution at 13 TeV to be used as reference forces us to base our projections on the ATLAS search for W' bosons in the dilepton channel [170]. As in the case of the NC projections (discussed in Sect. 2.3.3), theory predictions for the m_T distribution at high-mass are generated using the same selection and acceptance cuts as in [170] but now using an extended coverage in m_T. Further, we similarly restrict ourselves to events with either $m_{\ell\ell}$ or m_T greater than 500 GeV.

The pseudodata for the $m_{\ell\ell}$ (m_T) distribution at the HL-LHC is displayed in Fig. 3.8 (Fig. 3.9), with the highest energy bins reaching $m_{\ell\ell} \simeq 4$ TeV ($m_T \simeq 3.5$ TeV) for neutral-current (charged-current) scattering.

The percentage statistical and systematic uncertainties associated to the HL-LHC pseudo-data are displayed in the lower panels of Figs. 3.8 and 3.9, and are estimated using the same procedure as outlined in Sect. 2.3.3 (choosing to work with a five-fold

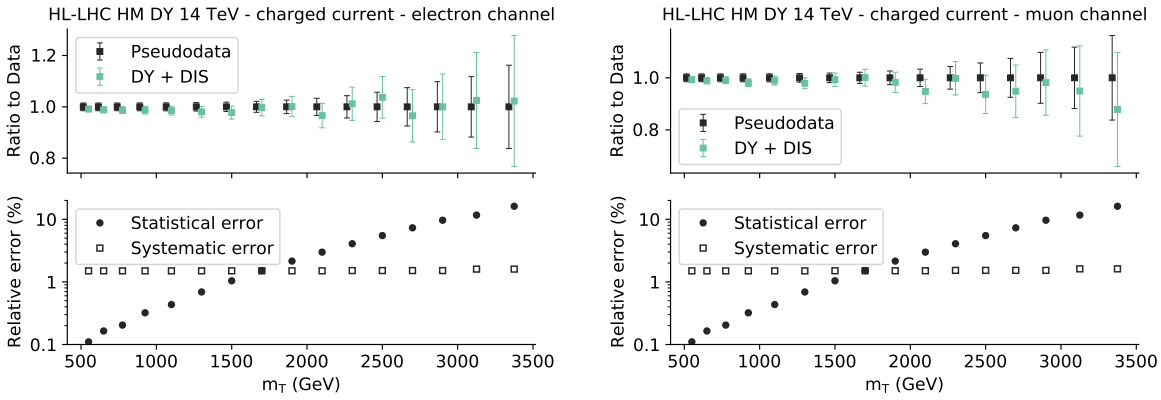


Figure 3.9: Same as Fig. 3.8 for charged-current Drell-Yan in bins of the transverse mass m_T.

reduction factor for the systematics, corresponding to the *optimistic* scenario there[3]). As in Sect. 2.3.3, the central values for the HL-LHC pseudodata are then generated by fluctuating the reference theory prediction by the expected total experimental uncertainty, namely:

$$\sigma_i^{\rm hllhc} \equiv \sigma_i^{\rm th}\left(1+\lambda\delta_{\mathcal{L}}^{\rm exp}+r_i\delta_{{\rm tot},i}^{\rm exp}\right)\,, \qquad i=1,\ldots,n_{\rm bin}\,, \tag{3.45}$$

where λ, r_i are univariate Gaussian random numbers, $\delta_{{\rm tot},i}^{\rm exp}$ is the total (relative) experimental uncertainty corresponding to this specific bin (excluding the luminosity and normalisation uncertainties), and $\delta_{\mathcal{L}}^{\rm exp}$ is the luminosity uncertainty, which is fully correlated amongst all the pseudo-data bins of the same experiment. Again, we take this luminosity uncertainty to be $\delta_{\mathcal{L}}^{\rm exp}=1.5\%$ for both ATLAS and CMS, as done in Ref. [119].

We have verified that, both at the pre- and post-fit levels, the fit quality to the HL-LHC pseudo-data satisfies $\chi^2/n_{\rm bin}\simeq 1$ in the case of the SM PDFs as expected.

3.6.2 Impact on PDF uncertainties

From Figs. 3.8 and 3.9, one can observe that the PDF uncertainties in the SM PDF baseline used to generate the pseudodata are either comparable or larger than the corresponding projected experimental uncertainties at the HL-LHC. Specifically, for the highest $m_{\ell\ell}$ bin of the NC distribution the PDF errors are twice the experimental ones, while in the CC case the associated PDF errors become clearly larger than the experimental ones starting at $m_T\simeq 2$ TeV. This comparison suggests that one should expect a significant uncertainty reduction once the HL-LHC pseudodata is included in the PDF fit.

[3]We note that the *conservative scenario* for the reduction of systematic errors, namely $f_{{\rm red},j}=0.5$, is not expected to qualitatively modify our results. The reason is that, as indicated by the bottom panels of Figs. 3.8 and 3.9, for the highest energy bins (which dominate the EFT sensitivity), specifically above $m_{\ell\ell}\approx 1.7$ TeV and $m_T\approx 1.5$ TeV, the measurement will be limited by statistical uncertainties.

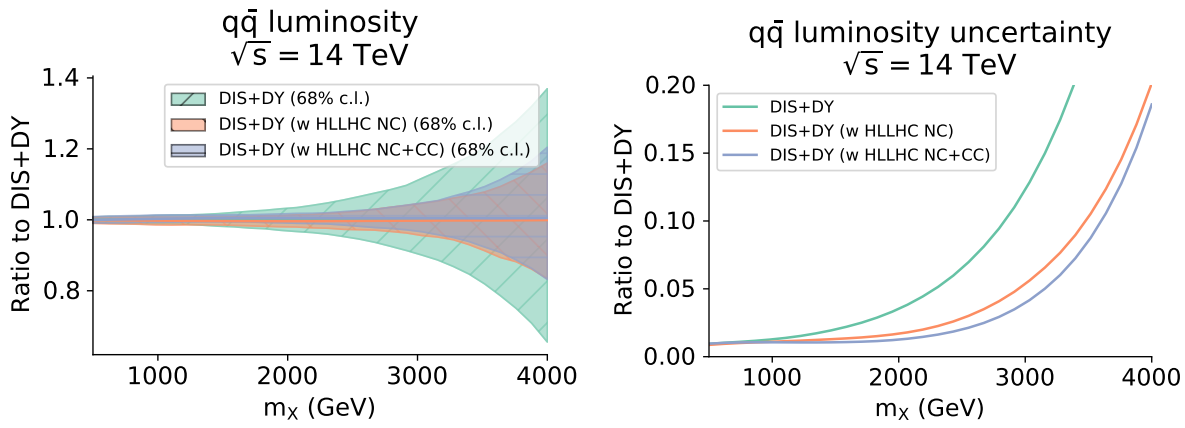


Figure 3.10: Impact of the HL-LHC pseudo-data on the quark-antiquark luminosity $\mathcal{L}_{q\bar{q}}$ of the SM PDF baseline fit as a function of m_X. Left: the luminosities $\mathcal{L}_{q\bar{q}}$ for the DIS+DY baseline and the corresponding fits including the HL-LHC pseudo-data, either only NC or also with CC cross sections, presented as a ratio to the central value of the former. Right: the relative PDF uncertainty in $\mathcal{L}_{q\bar{q}}$ (with the central value of the DIS+DY baseline as reference) for the same fits.

To validate this expectation, Fig. 3.10 displays the impact of the HL-LHC pseudo-data on the quark-antiquark luminosity $\mathcal{L}_{q\bar{q}}$ as a function of the final state invariant mass m_X at $\sqrt{s} = 14$ TeV. We compare $\mathcal{L}_{q\bar{q}}$ for the SM PDF baseline fit (DIS+DY) with the same quantity from the corresponding fits including the HL-LHC pseudo-data, either only NC or also with CC cross sections. The right panel displays the associated relative PDF uncertainties. We find a significant reduction of the PDF uncertainties affecting the quark-antiquark luminosity (and hence the Drell-Yan cross sections) in the high mass ($m_X \gtrsim 1$ TeV) region once the HL-LHC pseudo-data constraints are accounted for. For instance, at $m_X \gtrsim 2$ TeV, PDF uncertainties on $\mathcal{L}_{q\bar{q}}$ decrease from $\simeq 5\%$ in the baseline down to $\simeq 2.5\%$ ($\simeq 1.5\%$) once the NC (NC+CC) HL-LHC pseudo-data is included in the fit. The effect of the inclusion of HL-LHC projections becomes more dramatic as m_X increases. On the other hand, other partonic luminosities such as the quark-quark and gluon-gluon ones are essentially unaffected by the HL-LHC constraints. In terms of fit quality, the only noticeable effect is a mild improvement in the χ^2 of the high-mass DY datasets listed in Table 3.2.

3.6.3 PDF and EFT interplay at the HL-LHC

The finding that the projected HL-LHC pseudo-data has a significant impact on the quark-antiquark PDF luminosity, summarised in Fig. 3.10, suggests that the interplay between PDFs and EFT effects in the high-energy DY tails should become enhanced as compared to the results reported in the previous section. With this motivation, we first of all repeat

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$\hat{W} \times 10^5$ (68% CL)	$[-0.7, 0.5]$	$[-4.5, 6.9]$	1.3	850%
	$[-1.0, 0.9]$		1.3	500%
$\hat{W} \times 10^5$ (95% CL)	$[-1.0, 0.8]$	$[-8.1, 10.6]$	1.4	940%
	$[-1.4, 1.2]$		1.4	620%
$\hat{Y} \times 10^5$ (68% CL)	$[-1.8, 3.2]$	$[-6.4, 8.0]$	0.1	190%
	$[-3.7, 4.7]$		0.3	70%
$\hat{Y} \times 10^5$ (95% CL)	$[-3.4, 4.7]$	$[-11.1, 12.6]$	0.1	190%
	$[-5.3, 6.3]$		0.3	110%

Table 3.5: Same as Table 3.4 for the 68% CL and 95% CL marginalised bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the two-dimensional ($\hat{W}$,$\hat{Y}$) fits that include the HL-LHC pseudo-data for NC and CC Drell-Yan distributions. As in Table 3.4, for the SM PDFs we indicate the bounds obtained without (upper) and with (lower entry) PDF uncertainties accounted for.

the joint determination of PDFs and the $\hat{W}, \hat{Y}$ coefficients from the benchmark scenario presented in Sect. 3.5 now accounting for the constraints of the HL-LHC pseudo-data. An important difference in this case is that the inclusion of CC data lifts the flat direction in the $(\hat{W}, \hat{Y})$ plane, making a full two-dimensional fit possible.

For the simultaneous determination of PDFs and the $\hat{W}, \hat{Y}$ coefficients accounting for the constraints provided by the HL-LHC pseudo-data, we use 35 sampling values of $(\hat{W}, \hat{Y})$, 25 of which are equally spaced in either $\hat{W} \in (-1.6, 1.6) \times 10^{-5}$ or $\hat{Y} \in (-8, +8) \times 10^{-5}$ (hence in steps of $\Delta\hat{W} = 0.8 \times 10^{-6}$ and $\Delta\hat{Y} = 4 \times 10^{-6}$ respectively), and then 10 additional points along the diagonals. In order to assess the robustness of the results, we added 12 more sampling values, 8 further away from the origin and 4 more along the $\hat{W} = 0$ and $\hat{Y} = 0$ axes, and verified that the confidence level contours are stable upon their addition.

We find that the constraints on the $(\hat{W}, \hat{Y})$ parameters are completely dominated by the HL-LHC projections and that current data exhibit a much smaller pull, consistent with the findings of previous studies [120, 127]. Also, the $\chi^2_{\rm eftp}$ contour is more stable and requires less replicas if only the HL-LHC projections are included in the computation of the χ^2. The corresponding marginalised bounds on $\hat{W}$ and $\hat{Y}$ are reported in Table 3.5 using the same format as in Table 3.4.

From Table 3.5, one can observe how including high-mass data at the LHC both in a fit of PDFs and in a fit of SMEFT coefficients and neglecting the interplay between them could result in a significant underestimate of the uncertainties associated to the EFT parameters. Indeed, the marginalised 95% CL bound on the $\hat{W}$ ($\hat{Y}$) parameter becomes looser once SMEFT PDFs are consistently used, with a broadening, defined in Eq. (3.44),

	SM cons. PDFs	SMEFT PDFs	best-fit shift	broadening
$\hat{W} \times 10^5$ (68% CL)	$[-1.0, 0.0]$	$[-4.5, 6.9]$	1.7	1000%
	$[-4.0, 2.8]$		1.8	70%
$\hat{W} \times 10^5$ (95% CL)	$[-1.4, 0.4]$	$[-8.1, 10.6]$	1.8	940%
	$[-4.3, 3.1]$		1.9	150%
$\hat{Y} \times 10^5$ (68% CL)	$[2.1, 7.0]$	$[-6.4, 8.0]$	-3.7	190%
	$[-3.4, 11.2]$		-3.6	-1%
$\hat{Y} \times 10^5$ (95% CL)	$[0.5, 8.5]$	$[-11.1, 12.6]$	-3.7	200%
	$[-5.0, 13.7]$		-3.6	30%

Table 3.6: Same as Table 3.5 for the 68% and 95% CL marginalised bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the two-dimensional ($\hat{W}$,$\hat{Y}$) fits that include the HL-LHC pseudo-data for NC and CC Drell-Yan distributions. The input PDF set for the analysis done using fixed SM PDFs (corresponding to the results displayed in the column 'SM cons. PDFs') is a conservative PDF set that does not include any of the high-mass distributions or the HL-LHC projections nor the Run I and Run II high-mass dataset listed in Table 3.2. The limits obtained from the simultaneous fit of PDFs and Wilson coefficients (corresponding to the results displayed on the column 'SMEFT PDFs') are the same as those in Table 3.5.

of 500% (110%), even once PDF uncertainties are fully accounted for. This effect would have been even more marked if PDF uncertainties had not been accounted for in EFT fits based on SM PDFs, where the same broadening factors would be 940% and 190% respectively.

A further important question is whether the bounds obtained with SM PDFs appearing on the left column of Table 3.5 would become more comparable to those obtained from the simultaneous fit of PDFs and SMEFT coefficients, in case a conservative set of PDF was used in the analysis based on SM PDFs. To address this question, in Table 3.6 we display the bounds that are obtained using a PDF set that does not include any of the high-mass Drell-Yan sets (neither the HL-LHC projections nor the current datasets listed in Table. 3.2) and compare the bounds obtained using this set of PDFs to those obtained consistently using SMEFT PDFs. We observe that, once this set of conservative PDF is used as an input PDF set and the PDF uncertainty is included in the computation of the bounds, the latter increases as compared to the bounds in Table 3.5. As a result, the size of the bounds obtained by keeping fixed SM PDFs is closer to the size obtained from the simultaneous fits, although still slightly underestimated. At the same time, the shift in the best-fit becomes more marked.

Results are graphically displayed in Fig. 3.11, where the 95% confidence level contours in the ($\hat{W}$,$\hat{Y}$) plane obtained from the DIS+DY fits that include the high-mass Drell-Yan HL-LHC pseudo-data when using either SM PDFs, SM conservative PDFs or SMEFT

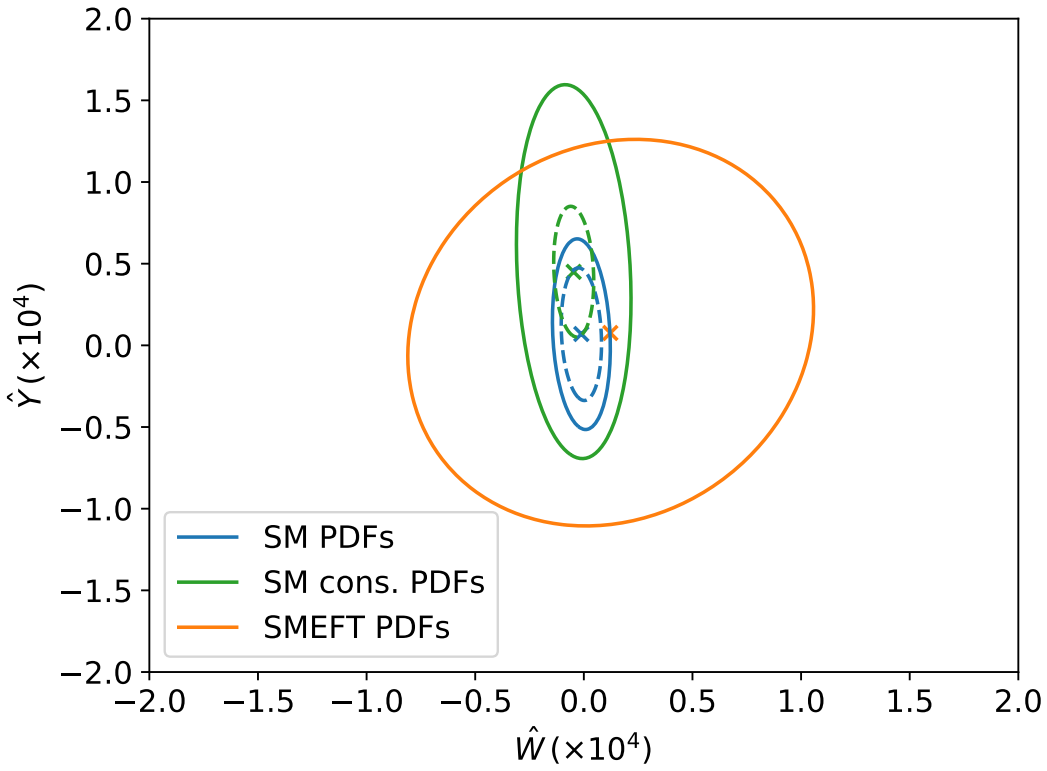


Figure 3.11: The 95% confidence level contours in the $(\hat{W},\hat{Y})$ plane obtained from the DIS+DY fits that include the high-mass Drell-Yan HL-LHC pseudo-data (both in the NC and CC channels) when using either SM PDFs (blue) or conservative SM PDFs (green). In both cases the ellipses are obtained by performing a parabolic fit to $\chi^2_{\rm smp}$ with fixed PDFs. PDF uncertainties are included in the solid lines and not included in the dashed lines. The results are compared to those obtained in a simultaneous fit, namely with SMEFT PDFs (orange). In this case, the parabolic fit is performed to $\chi^2_{\rm eftp}$ by varying simultaneously the Wilson Coefficients and the PDFs. The crosses indicate the best fits in the three cases discussed in the text.

PDFs are compared. All solid contours include PDF uncertainties, while the dashed contours that do not include PDF uncertainties are also indicated to visualise the impact of the inclusion of the PDF uncertainties.

To conclude, we should also emphasise that, while in this chapter we use pseudodata and hence the best-fit values are by construction unchanged, this would not necessarily be the case in the analysis of real data, where improper treatment of PDFs could result in a spurious EFT 'signal', or even missing a signal which is indeed present in the data. A detailed study aimed at a precise definition of 'conservative' PDFs is given in Chapter 5; a thorough comparison of the consistent simultaneous approach, versus the use of conservative PDF sets, will be of particular interest in cases of EFT manifestations of new physics.

The increased role that the interplay between PDFs and EFT coefficients will play

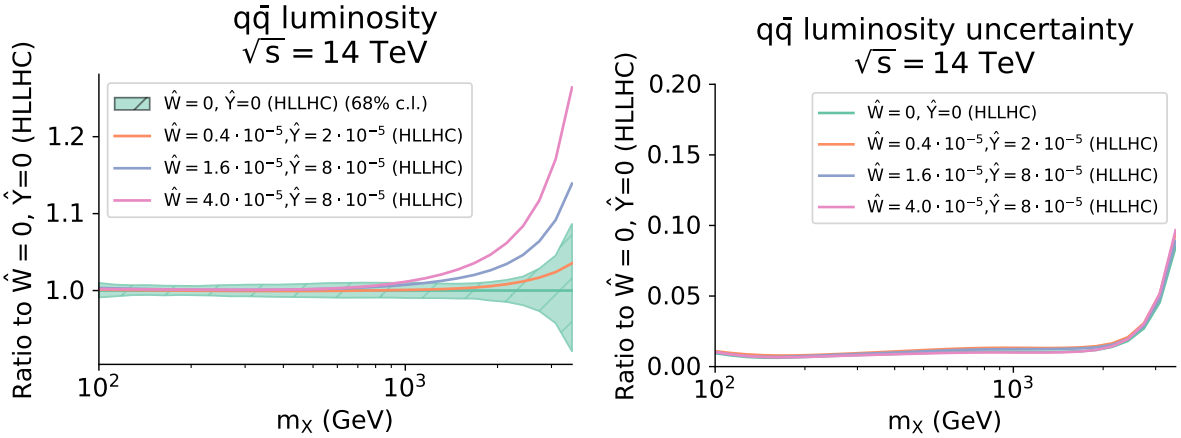


Figure 3.12: Same as Fig. 3.10, now comparing the quark-antiquark SM PDF luminosity in the fits including the HL-LHC pseudodata with those obtained in the SMEFT PDF fits for representative values of the $\hat{W}$ and $\hat{Y}$ parameters. The corresponding comparison in the case of fits to available Drell-Yan data was shown in Fig. 3.7.

at the HL-LHC can also be illustrated by comparing the expected behaviour of the quark-antiquark luminosity, displayed in Fig. 3.12, for the SMEFT PDFs corresponding to representative values of the $\hat{W}$ and $\hat{Y}$ parameters of the benchmark scenario as compared to the SM PDFs. Note that the corresponding comparison for $\mathcal{L}_{q\bar{q}}$ in the fits to available Drell-Yan data was displayed in Fig. 3.7. Indeed, the central value of the quark-antiquark luminosity for SMEFT PDFs corresponding to values of $(\hat{W}, \hat{Y})$ selected along the grid used to derive Fig. 3.12 changes greatly, well outside the one-sigma error band of the SM PDFs, while the PDF uncertainties themselves are unchanged. This change in central value of the large-x PDFs partially reabsorbs the effects in the partonic cross section induced by the SMEFT operators and leads to better χ^2 values as compared to those obtained with the SM PDFs.

Even neglecting SMEFT PDF effects, we note that our marginalised bounds on the $\hat{W}$ and $\hat{Y}$ coefficients from HL-LHC pseudo-data using SM PDFs turn out to be more stringent than those reported in [127] by around a factor of 4 for $\hat{W}$ and a factor 2 for $(\hat{Y})$. This is due to a combination of factors. First of all we use the 13 TeV measurements as reference to produce the HL projections. Furthermore we assume a total integrated luminosity of $\mathcal{L} = 6$ fb^{-1} (from the combination of ATLAS and CMS) rather than 3 fb^{-1} as well as a more optimistic scenario concerning the reduction of the experimental systematic uncertainties.

3.7 A first look at PDF 'contamination': injecting New Physics into the HL-LHC data

The study presented in Sect. 3.6 was based entirely on the assumption that data on NC and CC DY from the HL-LHC will be entirely consistent with the Standard Model. It is interesting to ask what happens if instead we inject the data with some New Physics, that is, if we base the generation of the HL-LHC pseudodata in Sect. 3.6 not on SM central values, but instead on SM central values multiplied by some non-unit K-factor combination of the $\hat{W}, \hat{Y}$ parameters. Such a study was performed in the conference proceedings given in Ref. [109], and is described here.

Let us suppose that we now generate the HL-LHC pseudodata in Sect. 3.6 with $\hat{W}, \hat{Y}$ fixed to the values $(\hat{W}, \hat{Y}) = (4, 8) \times 10^{-5}$, taking these values from within the 95% confidence intervals presented in Sect. 3.6. Performing the same analysis, and in particular recomputing the EFT bounds, we obtain the results given in Table 3.7.

	SM PDFs	SM cons. PDFs	SMEFT PDFs
$\hat{W} \times 10^5$ (95% CL)	$[-1.5, 1.2]$	$[3.1, 5.0]$	$[-5.3, 9.0]$
$\hat{Y} \times 10^5$ (95% CL)	$[-3.1, 8.7]$	$[5.8, 13.6]$	$[-0.2, 26.7]$

Table 3.7: We inject a spurious signal of new physics into the HL-LHC pseudodata, taking $(\hat{W}, \hat{Y}) = (4, 8) \times 10^{-5}$ as a benchmark. The table shows the 95% CL marginalised bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the two-dimensional $(\hat{W}, \hat{Y})$ fits that include this HL-LHC pseudodata. PDF uncertainties are accounted for.

We find that the fully simultaneous fit does a good job of detecting New Physics, with the bounds moving to the right relative to those in Sect. 3.6. In contrast, the fit using SM PDFs that have seen the SMEFT-affected data are unable to detect New Physics: the point $(\hat{W}, \hat{Y}) = (4, 8) \times 10^{-5}$ lies outside of the marginalised bounds at 95% CL shown in the leftmost column of Table 3.7. Finally we find that using conservative SM PDFs we are able to detect the New Physics, and the bounds are in fact tighter than those obtained using SMEFT PDFs. Our results suggest that a more careful study of conservative PDFs will be very important in the future, as PDF fits continue to include more and more data, some of which could be SMEFT-contaminated. In particular, it will be crucial for those performing SMEFT fits to know whether a fully simultaneous PDF-SMEFT fit is required, or whether they can reliably use conservative sets instead. A discussion of this point is given in the dedicated study presented in Chapter 5.

Chapter 4

Parton distributions in the SMEFT from the LHC Run II top dataset

[This chapter is based on Ref. [40], produced in collaboration with Zahari Kassabov, Maeve Madigan, Luca Mantani, Manuel Morales Alvarado, Juan Rojo and Maria Ubiali. My contributions to this study comprised: implementation of all new datasets included in this study in the NNPDF framework; production of all SM predictions for all datasets in this study, available as NLO grids suitable for PDF fitting and NNLO QCD K-factors; re-implementation and re-structuring of much of the SIMUNET *code, and extending it with new features and analysis tools (together with Manuel); running all SM PDF fits presented in this study, and a subset of the SMEFT-PDF fits; writing appendices to the study concerning fit quality and the pitfalls of the Monte Carlo replica method.]*

In Chapter 3, we introduced the SMEFT, and demonstrated that a simultaneous extraction of PDFs and SMEFT Wilson coefficients may be necessary for the high luminosity LHC. However, this study was limited in scope, fitting only two Wilson coefficients together with PDFs. In this chapter, we perform a much more comprehensive simultaneous extraction of PDFs, together with more than twenty SMEFT Wilson coefficients, focussing this time on those which affect processes involving top quark production.

Being the heaviest elementary particle known to date, with a mass around 185 times heavier than a proton, and the only fermion with an $\mathcal{O}(1)$ Yukawa coupling to the Higgs boson, the top quark has long been suspected to play a privileged role in potential new physics extensions beyond the Standard Model (BSM). For instance, radiative corrections involving top quarks are responsible for the so-called hierarchy problem of the SM, and the value of its mass m_t determines whether the vacuum state of our Universe is stable, metastable, or unstable [171, 172, 173]. For this reason, a comprehensive interpretation of the top sector within the framework of the SMEFT, and a precise discussion of the SMEFT-PDF interplay from the top sector, is an interesting avenue of study.

We begin in Sect. 4.1 with a review of our dataset, which combines standard data entering SM PDF fits with the broadest and most up-to-date subset of the Run II LHC top data ever considered. In Sect. 4.2, we continue to describe the SM theory predictions used for the top quark data given in Sect. 4.1. We additionally introduce the SMEFT operators which affect the top processes used in this chapter; further, we also describe the generation of the SMEFT theory predictions, and how these augment the existing SM theory calculations. In Sect. 4.4 we present the results of SM PDF fits using the dataset given in Sect. 4.1; these constitute the most comprehensive SM PDF fit of the top sector to date. Subsequently, we perform SMEFT-only fits to the top sector in Sect. 4.5, and compare these to previous SMEFT analyses. The key results of the analysis, namely the joint PDF-SMEFT fits, are presented in Sect. 4.6. Finally, we make comments on the use of our methodology for quadratic SMEFT fits in Sect. 4.7, which sets the stage for further discussion in Chapter 6.

4.1 The Run II top quark dataset

In this section, we describe the experimental data used in our subsequent analysis of PDFs and SMEFT Wilson coefficients. We begin in Sect. 4.1.1 by describing the datasets that we consider, with emphasis on the top quark production measurements. We proceed in Sect. 4.1.2 to use a modified version of the selection criteria defined in [31] to determine a maximally consistent dataset.

4.1.1 Experimental data

With the exception of the top quark measurements, the dataset used in this chapter for fitting the PDFs both in the SM-PDF and SMEFT-PDF cases overlaps with that of the NNPDF determination presented in Ref. [31]. In particular, the no-top variant of the NNPDF dataset consists of 4535 data points corresponding to a wide variety of processes in deep-inelastic lepton-proton scattering [174, 175, 176, 177, 178, 179, 180, 131, 181] and in hadronic proton-proton collisions [135, 133, 132, 136, 137, 138, 154, 182, 139, 141, 140, 155, 145, 183, 156, 148, 152, 151, 184, 153, 128, 185, 186, 187, 188, 189, 143, 147, 144, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199]; see [31] for more details.

Concerning the LHC top quark measurements considered in the present analysis, they partially overlap, but significantly extend, the top datasets included in global PDF fits such as NNPDF [31] as well as in SMEFT analyses of the top quark sector [200, 201]. Here we discuss in turn the different types of measurements to be included: inclusive $t\bar{t}$ cross sections and differential distributions; $t\bar{t}$ production asymmetries; the W-helicity fractions; associated top pair production with vector bosons and heavy quarks, including

$\bar{t}tZ$, $\bar{t}tW$, $\bar{t}t\gamma$, $\bar{t}t\bar{t}t$, $\bar{t}t\bar{b}b$; $t-$ and $s-$channel single top production; and associated single top and vector boson production.

Choice of kinematic distribution. Many of these measurements, in particular those targeting top quark pair production, are available differentially in several kinematic variables, as well as either absolute distributions, or distributions normalised to the fiducial cross-section. We must decide which of the available kinematic distributions associated to a given measurement should be included in the fit, and whether it is more advantageous to consider absolute or normalised distributions.

Regarding the former, we note that correlations between kinematic distributions are in general not available, and only one distribution at a time can be included without double-counting (one exception is the ATLAS $t\bar{t}$ lepton+jet measurement at $\sqrt{s} = 8$ TeV [202] where the full correlation matrix is provided). Therefore, wherever possible we include the top-pair invariant mass $m_{t\bar{t}}$ distributions with the rationale that these have enhanced sensitivity to SMEFT operators via energy-growing effects; they also provide direct information on the large-x PDFs. Otherwise, we consider the top or top-pair rapidity distributions, y_t and $y_{t\bar{t}}$ respectively, which also provide the sought-for information on the large-x PDFs; furthermore they benefit from moderate higher-order QCD and electroweak corrections [203].

Regarding the choice of absolute versus normalised distributions, we elect to use normalised distributions together with corresponding fiducial cross-sections throughout. Normalised distributions are typically more precise that their absolute counterparts, since experimental and theoretical errors partially cancel out when normalising. In addition, normalisation does not affect the PDF and EFT sensitivity of the measurement, provided the fiducial cross section measurements used for normalising are also accounted for. From the implementation point of view, since in a normalised measurement one bin is dependent on the others, we choose to exclude the bin with lowest $m_{t\bar{t}}$ value (the production threshold) to avoid losing sensitivity arising from the high-energy tails.

Inclusive $t\bar{t}$ production. A summary of the inclusive $t\bar{t}$ fiducial cross sections and differential distributions considered in this work is provided in Table 4.1. We indicate in each case the centre of mass energy $\sqrt{s}$, the final-state channel, the observable(s) used in the fit, the luminosity, and the number of data points $n_{\rm dat}$, together with the corresponding publication reference. In the last two columns, we indicate with a ✓ the datasets that are included for the first time here in a global PDF fit (specifically, those which are new with respect to NNPDF) and in a SMEFT interpretation (specifically, in comparison with the global fits of [200, 201]). The sets marked with brackets have already been included in previous studies, but are implemented here in a different manner (e.g. by changing spectra

or normalisation), as indicated in the table; more details are given in each paragraph of the section.

The ATLAS dataset comprises six total cross section measurements and five differential normalised cross section measurements. Concerning the latter, at 8 TeV we include three distributions from the dilepton and ℓ+jets channels. In the ℓ+jets channel, several kinematic distributions are available together with their correlations. Following the dataset selection analysis carried out in [31], we select to fit the y_t and $y_{t\bar{t}}$ distributions as done in the NNPDF baseline. At 13 TeV, we include the normalised cross sections differential in $m_{t\bar{t}}$ from the ℓ+jets and hadronic channels, with both measurements being considered for the first time here in the context of a PDF analysis.

Moving to CMS, in the inclusive $t\bar{t}$ category we consider five total cross section and four normalised differential cross section measurements. At $\sqrt{s} = 8$ TeV we include differential distributions in the ℓ+jets and dilepton channels, the latter being doubly differential in $y_{t\bar{t}}$ and $m_{t\bar{t}}$. The double-differential 8 TeV measurement is part of NNPDF, but there the $(y_t, m_{t\bar{t}})$ distribution was fitted instead. At 13 TeV, we include the $m_{t\bar{t}}$ distributions in the dilepton and ℓ+jets channels. In the latter case we include the single $m_{t\bar{t}}$ distribution rather than the double-differential one in $(m_{t\bar{t}}, y_{t\bar{t}})$, which is also available, since we find that the latter cannot be reproduced by the NNLO SM predictions. We present a dedicated analysis of the double-differential distribution in Sect. 4.5.3. As mentioned above, we will study the impact of our dataset selection choices by presenting variations of the baseline SM-PDF, fixed-PDF, and SMEFT-PDF analyses in the following sections.

$t\bar{t}$ asymmetry measurements. The $t\bar{t}$ production asymmetry at the LHC is defined as:

$$A_C = \frac{N(\Delta|y| > 0) - N(\Delta|y| < 0)}{N(\Delta|y| > 0) + N(\Delta|y| < 0)}\,, \tag{4.1}$$

with $N(P)$ being the number of events satisfying the kinematical condition P, and $\Delta|y| = |y_t| - |y_{\bar{t}}|$ is the difference between the absolute values of the top quark and anti-top quark rapidities. The asymmetry A_C can be measured either integrating over the fiducial phase space or differentially, for example binning in the invariant mass $m_{t\bar{t}}$. Measurements of A_C are particularly important in constraining certain SMEFT directions, in particular those associated to the two-light-two-heavy operators. However, they are unlikely to have an impact on PDF fitting due to their large experimental uncertainties; nevertheless, with the underlying motivation of a comprehensive SMEFT-PDF interpretation of top quark data, we consider here the A_C measurement as part of our baseline dataset, and hence study whether or not they also provide relevant PDF information. A summary of the asymmetry measurements included in this work is given in Table 4.2.

Exp.	$\sqrt{s}$ (TeV)	Channel	Observable	$\mathcal{L}$ (fb^{-1})	$n_{\rm dat}$	Ref.	New (PDF fits)	New (SMEFT fits)
ATLAS	7	dilepton	$\sigma(t\bar{t})$	4.6	1	[204]		(✓)
	8	dilepton	$\sigma(t\bar{t})$	20.3	1	[204]		(✓)
			$1/\sigma d\sigma/dm_{t\bar{t}}$	20.2	5	[205]	($y_{t\bar{t}} \to m_{t\bar{t}}$)	(absolute → ratio)
		ℓ+jets	$\sigma(t\bar{t})$	20.2	1	[206]	✓	(✓)
			$1/\sigma d\sigma/d\lvert y_t\rvert$	20.3	4	[202]		($m_{t\bar{t}}, p_t^T \to \lvert y_t\rvert, \lvert y_{t\bar{t}}\rvert$)
			$1/\sigma d\sigma/d\lvert y_{t\bar{t}}\rvert$	20.3	4	[202]		($m_{t\bar{t}}, p_t^T \to \lvert y_t\rvert, \lvert y_{t\bar{t}}\rvert$)
	13	dilepton	$\sigma(t\bar{t})$	36.1	1	[207]	✓	✓
		hadronic	$\sigma(t\bar{t})$	36.1	1	[208]	✓	✓
			$1/\sigma d^2\sigma/d\lvert y_{t\bar{t}}\rvert dm_{t\bar{t}}$	36.1	10	[208]	✓	✓
		ℓ+jets	$\sigma(t\bar{t})$	139	1	[209]		(✓)
			$1/\sigma d\sigma/dm_{t\bar{t}}$	36	8	[210]	✓	(absolute → ratio)
CMS	5	combination	$\sigma(t\bar{t})$	0.027	1	[211]		✓
	7	combination	$\sigma(t\bar{t})$	5.0	1	[212]		✓
	8	combination	$\sigma(t\bar{t})$	19.7	1	[212]		✓
		dilepton	$1/\sigma d^2\sigma/dy_{t\bar{t}}dm_{t\bar{t}}$	19.7	16	[213]	($m_{t\bar{t}}, y_t \to m_{t\bar{t}}, y_{t\bar{t}}$)	
		ℓ+jets	$1/\sigma d\sigma/dy_{t\bar{t}}$	19.7	9	[214]		
	13	dilepton	$\sigma(t\bar{t})$	43	1	[215]		(✓)
			$1/\sigma d\sigma/dm_{t\bar{t}}$	35.9	5	[216]		(absolute → ratio)
		ℓ+jets	$\sigma(t\bar{t})$	137	1	[217]	✓	✓
			$1/\sigma d\sigma/dm_{t\bar{t}}$	137	14	[217]	✓	✓

Table 4.1: The inclusive cross-sections and differential distributions for top quark pair production from ATLAS and CMS that we consider in this analysis. For each dataset, we indicate the experiment, the centre of mass energy $\sqrt{s}$, the final-state channel, the observable(s) used in the fit, the integrated luminosity $\mathcal{L}$ in inverse femtobarns, and the number of data points $n_{\rm dat}$, together with the corresponding publication reference. In the last two columns, we indicate with a ✓ the datasets that are included for the first time here in a global PDF fit and in a SMEFT interpretation, respectively. The sets marked with brackets have already been included in previous studies but here we account for their constraints in different manner (e.g. by changing spectra or normalisation), as indicated in the table and in the text description.

Experiment	$\sqrt{s}$(TeV)	Channel	Observable	$\mathcal{L}$ (fb^{-1})	n_{dat}	Ref.	New (PDF fits)	New (SMEFT fits)
ATLAS	8	dilepton	A_C	20.3	1	[218]	✓	
	13	ℓ+jets	A_C	139	5	[219]	✓	✓
CMS	8	dilepton	A_C	19.5	3	[220]	✓	
	13	ℓ+jets	A_C	138	3	[221]	✓	
ATLAS/CMS comb.	8	ℓ+jets	A_C	20	6	[222]	✓	

Table 4.2: Same as Table 4.1 for the $t\bar{t}$ asymmetry datasets.

W-helicity fractions. The W-helicity fractions F_0, F_L and F_R are PDF-independent observables sensitive to SMEFT corrections, and the dependence of the theory predictions with respect to the Wilson coefficients can be computed analytically. Since these W-helicity fractions are PDF-independent observables, to include them in the joint SMEFT-PDF analysis one has to extend the methodology presented in [103] to include in the fit datasets that either lack, or have negligible, PDF sensitivity and depend only on the EFT coefficients. We describe how this can be achieved within the SIMUNET framework in Sect. 4.3.

In Table 4.3 we list the LHC measurements of the W-helicity fractions considered in the current analysis. At $\sqrt{s} = 8$ TeV we include the combined ATLAS and CMS measurement from [223], while at 13 TeV we consider the ATLAS measurement of the W-helicities from [224], for the first time in a SMEFT fit.

Experiment	$\sqrt{s}$(TeV)	Observable	$\mathcal{L}$ (fb^{-1})	n_{dat}	Ref.	New (SMEFT fits)
ATLAS/CMS comb.	8	F_0, F_L	20	2	[223]	
ATLAS	13	F_0, F_L	139	2	[224]	✓

Table 4.3: Same as Table 4.1 for the W-helicity fraction measurements. These helicity fractions are PDF-independent and hence are only relevant in constraining the EFT coefficients.

Associated top quark pair production. The next class of observables that we discuss is associated $t\bar{t}$ production with a Z- or a W-boson (Table 4.4), a photon γ (Table 4.5), or a heavy quark pair ($t\bar{t}b\bar{b}$ or $t\bar{t}t\bar{t}$, Table 4.6). While measurements of $t\bar{t}V$ have been considered for SMEFT interpretations, we use them for the first time here in the context of a PDF determination. The rare processes $t\bar{t}\gamma$, $t\bar{t}b\bar{b}$, and $t\bar{t}t\bar{t}$ exhibit a very weak PDF sensitivity and hence in the present analysis their theory predictions are obtained using a fixed PDF, in the same manner as the W-helicity fractions in Table 4.3.

Concerning the $t\bar{t}Z$ and $t\bar{t}W$ data, from both ATLAS and CMS we use four fiducial cross section measurements at 8 TeV and 13 TeV, and one distribution differential in p_T^Z at

Exp.	$\sqrt{s}$ (TeV)	Observable	$\mathcal{L}$ (fb^{-1})	n_{dat}	Ref.	New (PDF fits)	New (SMEFT fits)
ATLAS	8	$\sigma(t\bar{t}Z)$	20.3	1	[225]	✓	
		$\sigma(t\bar{t}W)$	20.3	1	[225]	✓	
	13	$\sigma(t\bar{t}Z)$	36.1	1	[226]	✓	
		$1/\sigma d\sigma(t\bar{t}Z)/dp_T^Z$	139	6	[227]	✓	✓
		$\sigma(t\bar{t}W)$	36.1	1	[226]	✓	
CMS	8	$\sigma(t\bar{t}Z)$	19.5	1	[228]	✓	
		$\sigma(t\bar{t}W)$	19.5	1	[228]	✓	
	13	$\sigma(t\bar{t}Z)$	35.9	1	[229]	✓	
		$1/\sigma d\sigma(t\bar{t}Z)/dp_T(Z)$	77.5	3	[230]	✓	(absolute → ratio)
		$\sigma(t\bar{t}W)$	35.9	1	[229]	✓	

Table 4.4: Same as Table 4.1 for the measurements of top quark production in association with a vector boson.

Experiment	$\sqrt{s}$(TeV)	Observable	$\mathcal{L}$ (fb^{-1})	n_{dat}	Ref.	New (SMEFT fits)
ATLAS	8	$\sigma(t\bar{t}\gamma)$	20.2	1	[231]	
CMS	8	$\sigma(t\bar{t}\gamma)$	19.7	1	[232]	

Table 4.5: Same as Table 4.1 for $t\bar{t}$ production in association with a photon. Theory predictions for these observables adopt a fixed PDF.

13 TeV. These measurements are particularly interesting to probe SMEFT coefficients that modify the interactions between the top quark and the electroweak sector. For top-quark production associated with a photon, we include the fiducial cross-section measurements from ATLAS and CMS at 8 TeV; also available is a differential distribution at 13 TeV from ATLAS binned in the photon transverse momentum p_T^γ [240], but we exclude this from our analysis because of the difficulty in producing SMEFT predictions in the fiducial phase space (in the FITMAKER analysis, its inclusion is only approximate, and in SMEFiT this distribution is neglected entirely). Finally, we include fiducial measurements of $t\bar{t}b\bar{b}$ and $t\bar{t}t\bar{t}$ production at 13 TeV considering the data with highest luminosity for each available final state.

Inclusive single-top pair production. The inclusive single-top production data considered here and summarised in Table 4.7 comprises measurements of single-top production in the t-channel, which have previously been included in PDF fits [241, 31], as well as measurements of single-top production in the s-channel, which in the context of PDF studies have been implemented for the first time in this study. For t-channel

Experiment	$\sqrt{s}$(TeV)	Channel	Observable	$\mathcal{L}$ (fb^{-1})	n_{dat}	Ref.	New (SMEFT fits)
ATLAS	13	multi-lepton	$\sigma_{tot}(t\bar{t}t\bar{t})$	139	1	[233]	
		single-lepton	$\sigma_{tot}(t\bar{t}t\bar{t})$	139	1	[234]	✓
		ℓ+jets	$\sigma_{tot}(t\bar{t}b\bar{b})$	36.1	1	[235]	
CMS	13	multi-lepton	$\sigma_{tot}(t\bar{t}t\bar{t})$	137	1	[236]	
		single-lepton	$\sigma_{tot}(t\bar{t}t\bar{t})$	35.8	1	[237]	
		all-jet	$\sigma_{tot}(t\bar{t}b\bar{b})$	35.9	1	[238]	
		dilepton	$\sigma_{tot}(t\bar{t}b\bar{b})$	35.9	1	[239]	
		ℓ+jets	$\sigma_{tot}(t\bar{t}b\bar{b})$	35.9	1	[239]	✓

Table 4.6: Same as Table 4.1 for the measurements of $t\bar{t}$ production in association with a heavy quark pair. Theory predictions for these observables adopt a fixed PDF.

production, we consider the ATLAS and CMS top and anti-top fiducial cross sections at $\sqrt{s} = 7, 8$, and 13 TeV, as well as normalised y_t and $y_{\bar{t}}$ distributions at 7 and 8 TeV (ATLAS) and at 13 TeV (CMS). For s-channel production, no differential measurements are available and hence we consider fiducial cross-sections at 8 and 13 TeV from ATLAS and CMS.

Associated single top-quark production with weak bosons. Finally, Table 4.8 lists the measurements of associated single-top production with vector bosons included in our analysis. We consider fiducial cross-sections for tW production at 8 and 13 TeV from ATLAS and CMS in the dilepton and single-lepton final states, as well as the tZj fiducial cross-section at 13 TeV from ATLAS and CMS in the dilepton final state. In addition, kinematical distributions in tZj production from CMS at 13 TeV are considered for the first time here in an EFT fit. For these differential distributions, the measurement is presented binned in either p_T^Z or p_T^t; here, we take the former as default for consistency with the corresponding $t\bar{t}Z$ analysis.

Exp.	$\sqrt{s}$ (TeV)	Channel	Observable	$\mathcal{L}$ (fb^{-1})	n_{dat}	Ref.	New (PDF fits)	New (SMEFT fits)
ATLAS	7	t-channel	$\sigma_{tot}(t)$	4.59	1	[242]	(✓)	✓
			$\sigma_{tot}(\bar{t})$	4.59	1	[242]	(✓)	✓
			$1/\sigma d\sigma(tq)/dy_t$	4.59	3	[242]		✓
			$1/\sigma d\sigma(\bar{t}q)/dy_{\bar{t}}$	4.59	3	[242]		✓
	8	t-channel	$\sigma_{tot}(t)$	20.2	1	[243]	(✓)	✓
			$\sigma_{tot}(\bar{t})$	20.2	1	[243]	(✓)	✓
			$1/\sigma d\sigma(tq)/dy_t$	20.2	3	[243]		(✓)
			$1/\sigma d\sigma(\bar{t}q)/dy_{\bar{t}}$	20.2	3	[243]		(✓)
		s-channel	$\sigma_{tot}(t+\bar{t})$	20.3	1	[244]	✓	
	13	t-channel	$\sigma_{tot}(t)$	3.2	1	[245]	(✓)	
			$\sigma_{tot}(\bar{t})$	3.2	1	[245]	(✓)	
		s-channel	$\sigma_{tot}(t+\bar{t})$	139	1	[246]	✓	✓
CMS	7	t-channel	$\sigma_{tot}(t)+\sigma_{tot}(\bar{t})$	1.17, 1.56	1	[247]		✓
	8	t-channel	$\sigma_{tot}(t)$	19.7	1	[248]	(✓)	
			$\sigma_{tot}(\bar{t})$	19.7	1	[248]	(✓)	
		s-channel	$\sigma_{tot}(t+\bar{t})$	19.7	1	[249]	✓	
	13	t-channel	$\sigma_{tot}(t)$	2.2	1	[250]	(✓)	
			$\sigma_{tot}(\bar{t})$	2.2	1	[250]	(✓)	
			$1/\sigma d\sigma/d\lvert y^{(t)}\rvert$	35.9	4	[251]	✓	

Table 4.7: Same as Table 4.1 for the inclusive single-top production datasets.

Experiment	$\sqrt{s}$(TeV)	Channel	Observable	$\mathcal{L}$ (fb^{-1})	n_{dat}	Ref.	New (SMEFT fits)
ATLAS	8	dilepton	$\sigma_{tot}(tW)$	20.3	1	[252]	
		single-lepton	$\sigma_{tot}(tW)$	20.2	1	[253]	
	13	dilepton	$\sigma_{tot}(tW)$	3.2	1	[254]	
		dilepton	$\sigma_{fid}(tZj)$	139	1	[255]	
CMS	8	dilepton	$\sigma_{tot}(tW)$	12.2	1	[256]	
	13	dilepton	$\sigma_{tot}(tW)$	35.9	1	[257]	
		dilepton	$\sigma_{fid}(tZj)$	77.4	1	[258]	
		dilepton	$d\sigma_{fid}(tZj)/dp_T^t$	138	3	[259]	✓
		single-lepton	$\sigma_{tot}(tW)$	36	1	[260]	✓

Table 4.8: Same as Table 4.1 for single-top production in association with an electroweak bosons.

4.1.2 Dataset selection

The top quark production measurements listed in Tables 4.1-4.8 summarise all datasets that have been considered for the present analysis. In principle, however, some of these may need to be excluded from the baseline fit dataset to ensure that the baseline dataset is maximally consistent. Following the dataset selection procedure adopted in [31], here our baseline dataset is chosen to exclude datasets that may be either internally inconsistent or inconsistent with other measurements of the same process type. These inconsistencies can be of experimental origin, for instance due to unaccounted (or underestimated) systematic errors, or numerically unstable correlation models, as well as originating in theory, for example whenever a given process is affected by large missing higher-order perturbative uncertainties. Given that the ultimate goal of a global SMEFT analysis, such as the present one, is to unveil deviations from the SM, one should strive to deploy objective dataset selection criteria that exclude datasets affected by such inconsistencies, which are unrelated to BSM physics.

The first step is to run a global SM-PDF fit including all the datasets summarised in Tables 4.1-4.8 (and additionally a fit with the data summarised therein, but with the CMS measurement of the differential $t\bar{t}$ cross-section at 13 TeV in the ℓ+jets channel replaced with the double-differential measurement) and monitor in each case the following two statistical estimators:

- The total χ^2 per data point and the number of standard deviations n_σ by which the value of the χ^2 per data point differs from the median of the χ^2 distribution for a perfectly consistent dataset,

$$n_\sigma \equiv \frac{|\chi^2 - 1|}{\sigma_{\chi^2}} = \frac{|\chi^2 - 1|}{\sqrt{2/n_{\rm dat}}}\,, \tag{4.2}$$

  where the χ^2 in this case (and in the rest of the chapter unless specified) is the experimental χ^2 per data point, which is defined as

$$\chi^2 \equiv \chi^2_{\rm exp}/n_{\rm dat} \,=\, \frac{1}{n_{\rm dat}} \sum_{i,j=1}^{n_{\rm dat}} (D_i - T_i^0) \,\left(\mathrm{cov}_{\rm exp}^{-1}\right)_{ij} (D_j - T_j^0), \tag{4.3}$$

  where T_i^0 are the theoretical predictions computed with the central PDF replica, which is the average over the PDF replicas, and the experimental covariance matrix is the one defined for example in Eq. (3.1) of Ref. [261].

  Specifically, we single out for further examination datasets for which $n_\sigma \geq 3$ and $\chi^2 \geq 2$ per data point, where the poor description of the data is unlikely to be caused by a statistical fluctuation (note that these conditions relax those given in [31], which we hope gives the opportunity for the EFT to account for poor quality fits to data,

rather than immediately attributing poor fits to inconsistencies). The question is then to ascertain whether this poor χ^2 can be explained by non-zero EFT coefficients (and in such case it should be retained for the fit) or if instead there one can find other explanations, such as the ones mentioned above, that justify removing it from the baseline dataset.

- The metric Z defined in Ref. [262] which quantifies the stability of the χ^2 with respect to potential inaccuracies affecting the modelling of the experimental correlations. The calculation of Z relies exclusively on the experimental covariance matrix and is independent of the theory predictions. A large value of the stability metric Z corresponds to datasets with an unstable covariance matrix, in the sense that small changes in the values of the correlations between data points lead to large increases in the corresponding χ^2. Here we single out for further inspection datasets with $Z \geq 4$.

  As also described in [262], it is possible to regularise covariance matrices in a minimal manner to assess the impact of these numerical instabilities at the PDF or SMEFT fit level, and determine how they affect the resulting pre- and post-fit χ^2. To quantify whether datasets with large Z distort the fit results in a sizable manner, one can run fit variants applying this decorrelation procedure such that all datasets exhibit a value of the Z-metric below the threshold. We do not find it necessary to run such fits in this chapter.

In Tables 4.9 and 4.10 we list the outcome of such a global SM-PDF fit, where entries that lie above the corresponding threshold values for χ^2, n_σ, or Z are highlighted in boldface. In the last column, we indicate whether the dataset is flagged. For the flagged datasets, we carry out the following tests to ascertain whether it should be retained in the fit:

- For datasets with $n_\sigma > 3$ and $Z > 4$, we run a fit variant in which the covariance matrix is regularised. If, upon regularisation of the covariance matrix, the PDFs are stable and both the χ^2 per data point and the $|n_\sigma|$ decrease to a value below the respective thresholds of 2.0 and 3.0, we retain the dataset, else we exclude it.

- For datasets with $\chi^2 > 2.0$ and $n_\sigma > 3$ we carry out a fit variant where this dataset is given a very high weight. If in this high-weight fit variant the χ^2 and n_σ estimators improve to the point that their values lie below the thresholds without deteriorating the description of any of the other datasets included the dataset is kept, then the specific measurement is not inconsistent, it just does not have enough weight compared to the other datasets. See Ref. [31] for a detailed discussion on the size of the weight depending on the size of the dataset.

Experiment	$\sqrt{s}$ (TeV)	Observable, Channel	$n_{\rm dat}$	$\chi^2_{\rm exp}/n_{\rm dat}$	n_σ	Z	flag
ATLAS	7	$\sigma^{\rm tot}_{t\bar{t}}$, dilepton	1	**4.63**	2.57	1.00	no
		$\sigma^{\rm tot}(t)$, t-channel	1	0.76	-0.17	1.00	no
		$\sigma^{\rm tot}(\bar{t})$, t-channel	1	0.29	-0.50	1.00	no
		$1/\sigma d(tq)/dy_t$, t-channel	3	0.97	-0.04	1.28	no
		$1/\sigma d(\bar{t}q)/dy_{\bar{t}}$, t-channel	3	0.06	-1.15	1.39	no
	8	$\sigma^{\rm tot}_{t\bar{t}}$, dilepton	1	0.03	-0.69	1.00	no
		$1/\sigma d\sigma/dm_{t\bar{t}}$, dilepton	5	0.29	-1.12	1.61	no
		$\sigma^{\rm tot}_{t\bar{t}}$, ℓ+jets	1	0.28	-0.51	1.00	no
		$1/\sigma d\sigma/d\lvert y_t\rvert$, ℓ+jets	4	**2.86**	2.63	1.65	no
		$1/\sigma d\sigma/d\lvert y_{t\bar{t}}\rvert$, ℓ+jets	4	**3.37**	**3.35**	2.19	**yes (kept)**
		A_C, dilepton	1	0.67	-0.23	1.00	no
		$\sigma(t\bar{t}Z)$	1	0.23	-0.54	1.00	no
		$\sigma(t\bar{t}W)$	1	**2.44**	1.01	1.00	no
		$\sigma^{\rm tot}(t+\bar{t})$, s-channel	1	0.21	-0.56	1.00	no
		$\sigma^{\rm tot}(tW)$, dilepton	1	0.54	-0.33	1.00	no
		$\sigma^{\rm tot}(tW)$, single-lepton	1	0.71	-0.21	1.00	no
	13	$\sigma^{\rm tot}_{t\bar{t}}$, dilepton	1	1.41	0.29	1.00	no
		$\sigma^{\rm tot}_{t\bar{t}}$, hadronic	1	0.23	-0.54	1.000	no
		$1/\sigma d^2\sigma/d\lvert y_{t\bar{t}}\rvert dm_{t\bar{t}}$, hadronic	10	1.95	2.12	2.33	no
		$\sigma^{\rm tot}_{t\bar{t}}$, ℓ+jets	1	0.50	-0.35	1.00	no
		$1/\sigma d\sigma/dm_{t\bar{t}}$, ℓ+jets	8	1.83	1.66	**7.61**	no
		A_C, ℓ+jets	5	0.99	-0.02	1.41	no
		$\sigma(t\bar{t}Z)$	1	0.75	-0.18	1.00	no
		$1/\sigma d\sigma(t\bar{t}Z)/dp_T(Z)$	5	1.93	1.47	2.27	no
		$\sigma(t\bar{t}W)$	1	1.43	0.30	1.00	no
		$\sigma^{\rm tot}(t)$, t-channel	1	0.72	-0.20	1.00	no
		$\sigma^{\rm tot}(\bar{t})$, t-channel	1	0.39	-0.43	1.00	no
		$\sigma^{\rm tot}(t+\bar{t})$, s-channel	1	0.70	-0.21	1.00	no
		$\sigma^{\rm tot}(tW)$, dilepton	1	1.15	0.36	1.00	no

Table 4.9: For the ATLAS measurements that we consider in this work, we list the outcome of a global SM-PDF fit with all measurements listed in Tables 4.1-4.8 included. We display for each dataset the number of data points, the χ^2 per data point (Eq. (4.3), the number of standard deviations n_σ (Eq. (5.4)), and the stability metric Z defined in [262]. The entries that lie above the corresponding threshold values are highlighted in boldface, In the last column, we indicate whether the dataset is flagged and is either kept or removed. See text for more details.

Experiment	$\sqrt{s}$ (TeV)	Observable	$n_{\rm dat}$	$\chi^2_{\rm exp}/n_{\rm dat}$	n_σ	Z	flag
CMS	5	$\sigma^{\rm tot}_{t\bar{t}}$, combination	1	0.56	-0.31	1.00	no
	7	$\sigma^{\rm tot}_{t\bar{t}}$, combination	1	1.08	0.06	1.00	no
		$\sigma^{\rm tot}(t)+\sigma^{\rm tot}(\bar{t})$, t-channel	1	0.72	-0.20	1.00	no
	8	$\sigma^{\rm tot}_{t\bar{t}}$, combination	1	0.27	-0.52	1.00	no
		$1/\sigma d^2\sigma/dy_{t\bar{t}}dm_{t\bar{t}}$, dilepton	16	0.98	-0.06	2.33	no
		$1/\sigma d\sigma/dy_{t\bar{t}}$, ℓ+jets	9	1.15	0.31	1.63	no
		A_C, dilepton	3	0.05	-1.16	1.16	no
		$\sigma(t\bar{t}Z)$	1	0.47	-0.37	1.00	no
		$\sigma(t\bar{t}W)$	1	**2.27**	0.90	1.00	no
		$\sigma^{\rm tot}(t)$, t-channel	1	0.01	-0.70	1.00	no
		$\sigma^{\rm tot}(\bar{t})$, t-channel	1	0.09	-0.64	1.00	no
		$\sigma^{\rm tot}(t+\bar{t})$, s-channel	1	1.11	0.08	1.00	no
		$\sigma^{\rm tot}(tW)$, dilepton	1	0.38	-0.44	1.00	no
	13	$\sigma^{\rm tot}_{t\bar{t}}$, dilepton	1	0.06	-0.66	1.00	no
		$1/\sigma d\sigma/dm_{t\bar{t}}$, dilepton	5	**2.49**	2.36	1.61	no
		$\sigma^{\rm tot}_{t\bar{t}}$, ℓ+jets channel	1	0.22	-0.55	1.00	no
		$1/\sigma d\sigma/dm_{t\bar{t}}$, ℓ+jets	14	1.41	1.08	**4.57**	no
		$1/\sigma d\sigma/dm_{t\bar{t}}dy_t$, ℓ+jets	34	**6.43**	22.4	3.88	**yes (excl)**
		A_C, ℓ+jets	3	0.29	-0.87	1.00	no
		$\sigma(t\bar{t}Z)$	1	1.24	0.17	1.00	no
		$1/\sigma d\sigma(t\bar{t}Z)/dp_T(Z)$	3	0.59	-0.50	1.28	no
		$\sigma(t\bar{t}W)$	1	0.66	-0.24	1.00	no
		$\sigma^{\rm tot}(t)$, t-channel	1	0.88	-0.08	1.00	no
		$\sigma^{\rm tot}(\bar{t})$, t-channel	1	0.13	-0.62	1.00	no
		$1/\sigma d\sigma/d\lvert y^{(t)}\rvert$, t-channel	4	0.38	-0.88	1.70	no
		$\sigma^{\rm tot}(tW)$, dilepton	1	0.43	-0.40	1.00	no
		$\sigma^{\rm tot}(tW)$, single-lepton	1	**2.84**	1.30	1.00	no
ATLAS-CMS combination	8	A_C, ℓ+jets	6	0.602	-0.69	1.65	no

Table 4.10: Same as Table 4.9 for the CMS and combined ATLAS-CMS datasets. *Note carefully:* the row corresponding to the CMS doubly-differential distribution at 13 TeV in the ℓ+jets channel comes from a separate fit, where the corresponding 1D distribution is replaced by this dataset.

From the analysis of Tables 4.9 and 4.10, one finds that only two datasets in the inclusive top quark pair production (ℓ+jets final state) category are flagged as potentially problematic: the ATLAS $|y_{t\bar{t}}|$ distribution at 8 TeV and the CMS double-differential distributions in $m_{t\bar{t}}$ and y_t at 13 TeV. The first of these was already discussed in the NNPDF analysis [31]. It was observed that each of the four distributions measured by ATLAS and presented in Ref. [202] behave somewhat differently upon being given large weight. The χ^2 of all distributions significantly improves when given large weight. However, while for the top transverse momentum and top pair invariant mass distributions this improvement is accompanied by a rather significant deterioration of the global fit quality, in the case of the top and top pair rapidity distributions the global fit quality is very similar and only the description of jets deteriorates moderately. The rapidity distributions thus remain largely compatible with the rest of the dataset, hence they are kept.

Also shown in one row of Table 4.10 is the fit-quality information for the CMS double-differential distribution at 13 TeV in the ℓ+jets channel, from a separate fit wherein the CMS single differential distribution at 13 TeV in the ℓ+jets channel is replaced by this dataset. We find that the 2D set is described very poorly, with a $\chi^2 = 6.43$, corresponding to a 22σ deviation from the median of the χ^2 distribution for a perfectly consistent dataset. To investigate this further, we performed a weighted fit; however, we find that the χ^2 improves only moderately (from χ^2= 6.43 to $\chi^2 = 4.56$) and moreover the χ^2-statistic of the other datasets deteriorates significantly (with total χ^2 jumping from 1.20 to 1.28). The test indicates that the double-differential distribution is both incompatible with the rest of the data and also internally inconsistent given the standard PDF fit. Hence we exclude this dataset from our baseline and include instead the single-differential distribution in $m_{t\bar{t}}$, which is presented in the same publication [217] and is perfectly described in the baseline fit. To check whether the incompatibility we observe in the double-differential distribution can be cured by the inclusion of SMEFT corrections, we will run a devoted analysis presented in Sect. 4.5.3.

4.2 Theoretical predictions

In this section we describe the calculation settings adopted for the SM and SMEFT cross-sections used in the present analysis.

SM cross-sections. Theoretical predictions for SM cross-sections are evaluated at NNLO in perturbative QCD, whenever available, and at NLO otherwise. Predictions accurate to NLO QCD are obtained in terms of fast interpolation grids from MadGraph5_aMC@NLO [162, 263], interfaced to APPLgrid [163] or FastNLO [264, 265, 266] together with aMCfast [267] and APFELcomb [160]. Wherever available, NNLO QCD

corrections to matrix elements are implemented by multiplying the NLO predictions by bin-by-bin K-factors, see Sect. 2.3 in [116]. The top mass is set to $m_t = 172.5$ GeV for all processes considered.

In the case of inclusive $t\bar{t}$ cross sections and charge asymmetries, a dynamical scale choice of $\mu_R = \mu_F = H_T/4$ is adopted, where H_T denotes the sum of the transverse masses of the top and anti-top, following the recommendations of Ref. [268]. This scale choice ensures that the ratio of fixed order NNLO predictions to the NNLO+NNLL ones is minimised, allowing us to neglect theory uncertainties associated to missing higher orders beyond NNLO. To obtain the corresponding NNLO K-factors, we use the HighTEA public software [269], an event database for distributing and analysing the results of fixed order NNLO calculations for LHC processes. The NNLO PDF set used in the computation of these K-factors is either NNPDF3.1 or NNPDF4.0, depending on whether a given dataset was already included in the NNPDF4.0 global fit or not, respectively.

For associated $t\bar{t}$ and W or Z production, dedicated fast NLO grids have been generated. Factorisation and renormalisation scales are fixed to $\mu_F = \mu_R = m_t + \frac{1}{2}m_V$, where $m_V = m_W, m_Z$ is the mass of the associated weak boson, as appropriate. This scale choice follows the recommendation of Ref. [270] and minimises the ratio of the NLO+NLL over the fixed-order NLO prediction. We supplement the predictions for the total cross section for associated W and Z-production at 13 TeV with NLO+NNLL QCD K-factors taken from Table 1 of [270]. On the other hand, the $t\bar{t}\gamma$, $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ data are implemented as PDF independent observables, and the corresponding theory predictions are taken directly from the relevant experimental papers in each case.

The evaluation of theoretical predictions for single top production follows [241]. Fast NLO interpolation grids are generated for both s- and t-channel single top-quark and top-antiquark datasets in the 5-flavour scheme, with fixed factorisation and renormalisation scales set to m_t. Furthermore, for the t-channel production we include the NNLO QCD corrections to both total and differential cross sections [271]. When the top decay is calculated, it is done in the narrow-width approximation, under which the QCD corrections to the top-(anti)quark production and the decay are factorisable and the full QCD corrections are approximated by the vertex corrections.

SMEFT cross-sections. SMEFT corrections to SM processes are computed both at LO and at NLO in QCD, and both at the linear and the quadratic level in the EFT expansion. Flavour assumptions follow the LHC TOP WG prescription of [272] which were also used in the recent SMEFiT analysis [201]. The flavour symmetry group is given by $U(3)_l \times U(3)_e \times U(3)_d \times U(2)_u \times U(2)_q$, i.e. we single out operators that contain top quarks (right-handed t and $SU(2)$ doublet Q). This also means that one works in a five-flavour scheme in which the only massive fermion in the theory is the top. As far as

the electroweak input scheme is concerned, we work in the m_W-scheme, meaning that the 4 electroweak inputs are $\{m_W, G_F, m_h, m_Z\}$. In particular, the electric charge e becomes a dependent parameter and is shifted by the effects of higher-dimensional operators.

At dimension-six, SMEFT operators modify the SM Lagrangian as:

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_{n=1}^{N} \frac{c_n}{\Lambda^2} \mathcal{O}_n \,, \tag{4.4}$$

where Λ is the UV-cutoff energy scale, $\{\mathcal{O}_n\}$ are dimension-six operators, and $\{c_n\}$ are Wilson coefficients. The 25 dimension-six operators considered in this chapter are listed in Table 4.11 in the Warsaw basis [114].[1] The upper part in Table 4.11 defines the relevant two-fermion operators modifying the interactions of the third-generation quarks. We also indicate the notation used for the associated Wilson coefficients; those in brackets are not degrees of freedom (DoF) entering the fit, and instead the two additional DoF defined in the middle table are used. The bottom table defines the four-fermion DoF entering the fit, expressed in terms of the corresponding four-fermion Wilson coefficients associated to dimension-six SMEFT operators in the Warsaw basis.

For hadronic data, i.e. for proton-proton collisions, which are the only data affected by the SMEFT in this study, the linear effect of the n-th SMEFT operator on a theoretical prediction can be quantified by:

$$R^{(n)}_{\text{SMEFT}} \equiv \left(\mathcal{L}^{\text{NNLO}}_{ij} \otimes d\widehat{\sigma}^{(n)}_{ij,\text{SMEFT}}\right) / \left(\mathcal{L}^{\text{NNLO}}_{ij} \otimes d\widehat{\sigma}_{ij,\text{SM}}\right) \,, \quad n = 1, \ldots, N, \tag{4.5}$$

where i, j are parton indices, $\mathcal{L}^{\text{NNLO}}_{ij}$ is the NNLO partonic luminosity defined as

$$\mathcal{L}_{ij}(\tau, M_X) = \int_\tau^1 \frac{dx}{x} \, f_i(x, M_X^2) f_j(\tau/x, M_X^2) \,, \quad \tau = M_X^2/s, \tag{4.6}$$

$d\widehat{\sigma}_{ij,\text{SM}}$ the bin-by-bin partonic SM cross section, and $d\widehat{\sigma}^{(n)}_{ij,\text{SMEFT}}$ the corresponding partonic cross section associated to the interference between $\mathcal{O}_n$ and the SM amplitude $\mathcal{A}_{\text{SM}}$ when setting $c_n = 1$. This value of c_n is only used to initialise the potential contributions of the SMEFT operator; the effective values of the Wilson coefficient are found after the fit is performed. Quadratic effects of the interference between the n-th and m-th SMEFT operators can be evaluated as

$$R^{(n,m)}_{\text{SMEFT}} \equiv \left(\mathcal{L}^{\text{NNLO}}_{ij} \otimes d\widehat{\sigma}^{(n,m)}_{ij,\text{SMEFT}}\right) / \left(\mathcal{L}^{\text{NNLO}}_{ij} \otimes d\widehat{\sigma}_{ij,\text{SM}}\right) \,, \quad n, m = 1, \ldots, N, \tag{4.7}$$

with the bin-by-bin partonic cross section $d\widehat{\sigma}^{(n,m)}_{ij,\text{SMEFT}}$ now being evaluated from the squared amplitude $\text{Re}(\mathcal{A}_n \mathcal{A}_m^* + \mathcal{A}_n^* \mathcal{A}_m)$ associated to the operators $\mathcal{O}_n$ and $\mathcal{O}_m$ when $c_n = c_m = 1$.

[1] Note that in this chapter, we neglect renormalisation group effects on the Wilson coefficients [273].

Operator	Coefficient	Definition
$\mathcal{O}_{\varphi Q}^{(1)}$	$(c_{\varphi Q}^{(1)})$	$i\big(\varphi^\dagger \overset{\leftrightarrow}{D}_\mu \varphi\big)\big(\bar{Q}\,\gamma^\mu\, Q\big)$
$\mathcal{O}_{\varphi Q}^{(3)}$	$c_{\varphi Q}^{(3)}$	$i\big(\varphi^\dagger \overset{\leftrightarrow}{D}_\mu \tau_I \varphi\big)\big(\bar{Q}\,\gamma^\mu\, \tau^I Q\big)$
$\mathcal{O}_{\varphi t}$	$c_{\varphi t}$	$i\big(\varphi^\dagger \overset{\leftrightarrow}{D}_\mu \varphi\big)\big(\bar{t}\,\gamma^\mu\, t\big)$
$\mathcal{O}_{tW}$	c_{tW}	$i\big(\bar{Q}\tau^{\mu\nu}\,\tau_I\, t\big)\,\tilde{\varphi}\,W_{\mu\nu}^I + \text{h.c.}$
$\mathcal{O}_{tB}$	(c_{tB})	$i\big(\bar{Q}\tau^{\mu\nu}\, t\big)\,\tilde{\varphi}\,B_{\mu\nu} + \text{h.c.}$
$\mathcal{O}_{tG}$	c_{tG}	$i\,\big(\bar{Q}\tau^{\mu\nu}\,T_A\, t\big)\,\tilde{\varphi}\,G_{\mu\nu}^A + \text{h.c.}$

DoF	Definition
$c_{\varphi Q}^{(-)}$	$c_{\varphi Q}^{(1)} - c_{\varphi Q}^{(3)}$
c_{tZ}	$-\sin\theta_W c_{tB} + \cos\theta_W c_{tW}$

DoF	Definition (Warsaw basis)
c_{QQ}^1	$2c_{qq}^{1(3333)} - \frac{2}{3}c_{qq}^{3(3333)}$
c_{QQ}^8	$8c_{qq}^{3(3333)}$
c_{Qt}^1	$c_{qu}^{1(3333)}$
c_{Qt}^8	$c_{qu}^{8(3333)}$
c_{tt}^1	$c_{uu}^{(3333)}$
$c_{Qq}^{1,8}$	$c_{qq}^{1(i33i)} + 3c_{qq}^{3(i33i)}$
$c_{Qq}^{1,1}$	$c_{qq}^{1(ii33)} + \frac{1}{6}c_{qq}^{1(i33i)} + \frac{1}{2}c_{qq}^{3(i33i)}$
$c_{Qq}^{3,8}$	$c_{qq}^{1(i33i)} - c_{qq}^{3(i33i)}$
$c_{Qq}^{3,1}$	$c_{qq}^{3(ii33)} + \frac{1}{6}(c_{qq}^{1(i33i)} - c_{qq}^{3(i33i)})$
c_{tq}^8	$c_{qu}^{8(ii33)}$
c_{tq}^1	$c_{qu}^{1(ii33)}$
c_{tu}^8	$2c_{uu}^{(i33i)}$
c_{tu}^1	$c_{uu}^{(ii33)} + \frac{1}{3}c_{uu}^{(i33i)}$
c_{Qu}^8	$c_{qu}^{8(33ii)}$
c_{Qu}^1	$c_{qu}^{1(33ii)}$
c_{td}^8	$c_{ud}^{8(33jj)}$
c_{td}^1	$c_{ud}^{1(33jj)}$
c_{Qd}^8	$c_{qd}^{8(33jj)}$
c_{Qd}^1	$c_{qd}^{1(33jj)}$

Table 4.11: Upper table: definition of the two-fermion dimension-six SMEFT operators relevant for this analysis. These operators modify the interactions of the third-generation quarks. We also indicate the notation for the associated Wilson coefficients; those in brackets are not degrees of freedom entering the fit. Middle table: the two additional degrees of freedom used in the fit involving two-fermion operators, defined in terms of the coefficients of the upper table. Bottom table: the four-fermion degrees of freedom considered here, expressed in terms of the corresponding four-fermion Wilson coefficients of dimension-six SMEFT operators in the Warsaw basis. Throughout, repeated indices indicate summation.

The computation of the SMEFT contributions is performed numerically with the FeynRules [274] model SMEFTatNLO [275], which allows one to include NLO QCD corrections to the observables. The obtained cross sections are then combined in so-called *BSM factors* by taking the ratio with the respective SM cross sections, in order to produce $R^{(n)}_{\rm SMEFT}$ and $R^{(n,m)}_{\rm SMEFT}$, respectively the linear and quadratic corrections.

With these considerations, we can account for SMEFT effects in our theoretical predictions by mapping the SM prediction $T^{\rm SM}$ to:

$$T = T^{\rm SM} \times K(\{c_n\}) \,, \tag{4.8}$$

with:

$$K(\{c_n\}) = 1 + \sum_{n=1}^{N} c_n R^{(n)}_{\rm SMEFT} + \sum_{1\leq n\leq m\leq N} c_{nm} R^{(n,m)}_{\rm SMEFT} \,, \tag{4.9}$$

with $c_{nm} = c_n c_m$. Eq. (4.8) is at the centre of the SIMUnet methodology, which we discuss in Sect. 4.3.

4.3 Fitting methodology

In this chapter, the joint determination of the PDFs and the EFT coefficients is carried out using the SIMUnet methodology, first presented in [103], which is substantially extended in this work. The core idea of SIMUnet is to incorporate the Wilson coefficients into the optimisation problem that enters the PDF determination, by accounting explicitly for their dependence in the theoretical predictions used to fit the PDFs. Specifically, the neural network model used in the SM-PDF fits of NNPDF is augmented with an additional layer, which encodes the dependence of the theory predictions entering the fit on the Wilson coefficients.

In this section, first we provide an overview of the SIMUnet methodology, highlighting the new features that have been implemented for the present study.

4.3.1 SIMUnet overview

The SIMUnet [103] methodology extends the NNPDF framework [31, 276] to account for the EFT dependence (or, in principle, any parametric dependence) of the theory cross-sections entering the PDF determination. This is achieved by adding an extra layer to the NNPDF neural network to encapsulate the dependence of the theory predictions on the EFT coefficients, including the free parameters in the general optimisation procedure. This results in a simultaneous fit of the PDF as well as EFT coefficients to the input data. As in the NNPDF methodology, the error uncertainty estimation makes use of the Monte Carlo replica method, which yields an uncertainty estimate on both PDF and

EFT parameters. We discuss the limitations of this method in Sect. 4.7, and further in Chapter 6.

The SM theoretical observables are encoded using interpolation grids, known as `FK`-tables [277, 278, 160], which capture the contribution of both the DGLAP evolution and the hard-scattering matrix elements and interface it with the initial-scale PDFs in a fast and efficient way.

The simultaneous fit is represented as a neural network using the `Tensorflow` [279] and `Keras` [280] libraries. The architecture is schematically represented in Fig. 4.1. Trainable weights are represented by solid arrows, and non-trainable weights by dashed arrows. Through a forward pass across the network, the inputs (x-Bjorken and its logarithm) proceed through hidden layers to output the eight fitted PDFs at the initial parametrisation scale Q_0. For each of the experimental observables entering the fit, these PDFs are then combined into a partonic luminosity $\mathcal{L}^{(0)}$ at Q_0, which is convolved with the precomputed `FK`-tables Σ to obtain the SM theoretical prediction $\mathcal{T}^{\rm SM}$. Subsequently, the effects of the N EFT coefficients $\boldsymbol{c} = (c_1, \ldots, c_N)$, associated to the operator basis considered, are accounted for by means of an extra layer, resulting in the final prediction for the observable $\mathcal{T}$ entering the SMEFT-PDF fit. The SIMUNET code allows for both linear and quadratic dependence on the EFT coefficients. In linear EFT fits, the last layer consists of N trainable weights to account for each Wilson coefficient. In quadratic EFT fits, in addition to the N trainable weights, a set of $N(N+1)/2$ non-trainable parameters, which are functions of the trainable weights, is included to account for all diagonal and non-diagonal contributions of EFT-EFT interference to the cross-sections. The results obtained with the quadratic functionality of SIMUNET are, however, not displayed in this chapter, for the reasons explained in Sect. 4.7. The PDF parameters $\boldsymbol{\theta}$ and the EFT coefficients $\boldsymbol{c}$ entering the evaluation of the SMEFT observable in Fig. 4.1 are then determined simultaneously from the minimisation of the fit figure of merit (the loss function described in the introductory chapter, namely the t_0 modified χ^2 with positivity and integrability penalty terms).

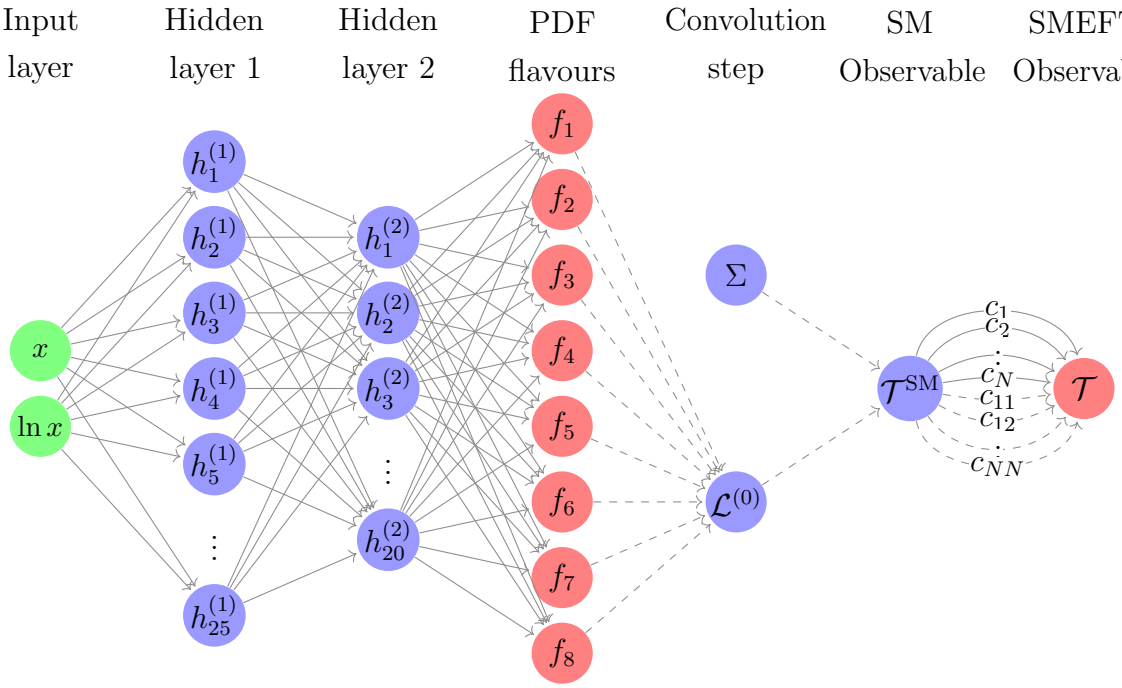


Figure 4.1: Schematic representation of the SIMUNET architecture for a general observable. Trainable weights are represented by solid arrows, and non-trainable weights by dashed arrows. Through a forward pass across the network, the inputs (x-Bjorken and its logarithm, in green) proceed through 2 hidden layers (in blue) to output the PDFs $f_1, \cdots, f_8$ (in red) at the initial parametrisation scale Q_0. For each of the experimental observables entering the fit, these PDFs are combined into a partonic luminosity $\mathcal{L}^{(0)}$ at Q_0, which is then convolved with precomputed `FK`-tables Σ to obtain the SM theoretical prediction $\mathcal{T}^{\rm SM}$. Subsequently, the effects of the EFT coefficients c_i are accounted for by means of an extra layer. In linear EFT fits this layer simplifies to just N trainable weights to account for each coefficient, and in quadratic EFT fits a set of $N(N+1)/2$ non-trainable weights has to be added to account for the EFT-EFT interference. The forward-pass of this layer results in the final prediction for the observable $\mathcal{T}$ entering the SMEFT-PDF fit. By setting the weights in the EFT layer to zero, one recovers the SM-PDF case. By freezing the PDF-related weights in the network architecture, one can carry out a fixed-PDF EFT determination or include in the joint SMEFT-PDF fit observables whose PDF dependence can be neglected.

The SIMUNET architecture can be minimally modified to deal with the fixed-PDF case, in which only the EFT coefficients are treated as free parameters in the optimisation process. This can be achieved by freezing the PDF-related weights in the network architecture to the values obtained in some previous fit, for example a SM-PDF determination based on NNPDF. In this manner, SIMUNET can also be used to carry out traditional EFT fits where the PDF dependence of the theory predictions is neglected. Furthermore, for PDF-independent observables, computing an `FK`-table Σ is not required and the SM cross-section $\mathcal{T}^{\rm SM}$ can be evaluated separately and stored to be used in the fit.

As illustrated in Fig. 4.1, within the SIMUNET framework a single neural network encapsulates both the PDF and the EFT dependence of physical observables, with the

corresponding parameters being simultaneously constrained from the experimental data included in the fit. Specifically, we denote the prediction of the neural network as:

$$\mathcal{T} = \mathcal{T}(\boldsymbol{\theta}) = (T_1(\boldsymbol{\theta}), \ldots, T_n(\boldsymbol{\theta})) \ , \tag{4.10}$$

with $n = n_{\rm dat}$ and $\hat{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \boldsymbol{c})$, where $\boldsymbol{\theta}$ and $\boldsymbol{c} = (c_1, \ldots, c_N)$ represent the weights associated to the PDF nodes of the network, and to the N Wilson coefficients from the operator basis, respectively. The uncertainty estimation uses the Monte Carlo replica method, where a large number $N_{\rm rep}$ of replicas $D^{(k)} = \left(D_1^{(k)}, \ldots, D_n^{(k)}\right)$ of the experimental measurements $D = (D_1, \ldots, D_n)$ are sampled from the distribution of experimental uncertainties with $k = 1, \ldots, N_{\rm rep}$. The optimal values for the fit parameters $\hat{\boldsymbol{\theta}}^{(k)}$ associated to each replica are obtained by means of a Stochastic Gradient Descent (SGD) algorithm that minimises the corresponding figure of merit:

$$E_{\rm tot}^{(k)}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{n_{\rm dat}} \sum_{i,j=1}^{n_{\rm dat}} \left(D_i^{(k)} - T_i(\hat{\boldsymbol{\theta}})\right) \left(\mathrm{cov}_{t^0}^{-1}\right)_{ij} \left(D_j^{(k)} - T_j(\hat{\boldsymbol{\theta}})\right) \ , \tag{4.11}$$

where the covariance matrix in Eq. (4.11) is the t_0 covariance matrix, which is constructed from all sources of statistical and systematic uncertainties that are made available by the experiments with correlated multiplicative uncertainties treated via the t_0 prescription [35] in the fit to avoid fitting bias associated with multiplicative uncertainties, as described in the introductory chapter. This loss is also augmented by positivity and integrability penalty terms, also described in the introductory chapter.

Once Eq. (4.11) is minimised for each replica, subject to the usual cross-validation stopping, one ends up with a sample of best-fit values for both the EFT coefficients and the PDF parameters:

$$\hat{\boldsymbol{\theta}}^{(k)} = \left(\boldsymbol{\theta}^{(k)}, \mathbf{c}^{(k)}\right) = \arg\min_{\hat{\boldsymbol{\theta}}} E_{\rm tot}^{(k)}\left(\hat{\boldsymbol{\theta}}\right) \ , \qquad k = 1, \ldots, N_{\rm rep} \, , \tag{4.12}$$

from which one can evaluate statistical properties such as averages, variances, higher moments, or confidence level intervals. For example, we could compute the mean of the sample via:

$$c_\ell^* = \left\langle c_\ell^{(k)} \right\rangle_{\rm rep} = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} c_\ell^{(k)} \ , \tag{4.13}$$

though it is worth noting that it can be more informative to compute the mode or median of the distribution, depending on its shape. Note that, in this methodology, the Monte Carlo error propagation automatically propagates the PDF uncertainty to the distribution of the best-fit values of the EFT coefficients.[2] Hence the variance on the EFT coefficients

[2]At least, it has been traditionally thought to do so. This is not in fact the case - see Sect. 4.7 and

reflects not only the experimental uncertainty of the data included in the fit, but also the functional uncertainty associated with the PDFs.

As we discuss below, the current implementation of the SIMUNET methodology also allows performing fixed-PDF fits, where only the Wilson coefficients are optimised. This is done by freezing the weights of the PDF part of the neural network during the minimisation of the loss function (4.11) from some other previous fit, $\boldsymbol{\theta}^{(k)} = \widetilde{\boldsymbol{\theta}}^{(k)}$, such that Eq. (4.12) reduces to

$$\hat{\boldsymbol{\theta}}^{(k)} = \left(\tilde{\boldsymbol{\theta}}^{(k)}, \mathbf{c}^{(k)}\right) = \arg\min_{\mathbf{c}} E_{\rm tot}^{(k)}\left(\tilde{\boldsymbol{\theta}}^{(k)}, \boldsymbol{c}\right) , \qquad k = 1, \ldots, N_{\rm rep} \, . \tag{4.14}$$

In this limit, SIMUNET reduces to a fixed-PDF EFT fit such as the MCfit variant of SMEFIT [43]. Likewise, by setting to zero the EFT coefficients,

$$\hat{\boldsymbol{\theta}}^{(k)} = \left(\boldsymbol{\theta}^{(k)}, \mathbf{c}^{(k)} = \mathbf{0}\right) = \arg\min_{\boldsymbol{\theta}} E_{\rm tot}^{(k)}\left(\boldsymbol{\theta}^{(k)}, \mathbf{c}^{(k)} = \mathbf{0}\right) , \qquad k = 1, \ldots, N_{\rm rep} \, , \tag{4.15}$$

one recovers the same PDF weights $\boldsymbol{\theta}^{(k)}$ as in NNPDF, or those of the SM-PDF fit being used as baseline in the analysis.

An important *caveat* here is that, while in the SIMUNET methodology the PDF uncertainty is propagated to the posterior distribution of the EFT coefficients via the Monte Carlo replica method, in the MCfit variant of the SMEFITmethodology the fit of the EFT only considers the central PDF member (which in the NNPDF case corresponds to the average of the PDF replicas) for all $N_{\rm rep}$ replicas, and the PDF uncertainty is propagated to the EFT coefficients by utilising an additional covariance matrix (both in the fit of the EFT coefficients and in the generation of the Monte Carlo replicas of the experimental data) that is added to t_0 covariance matrix. Namely,

$$\mathrm{cov}_{\rm exp+th} = \mathrm{cov}_{t^0} + \mathrm{cov}_{\rm th}, \tag{4.16}$$

where $\mathrm{cov}_{\rm th}$ includes the PDF contribution [201, 107], computed as

$$\left(\mathrm{cov}_{\rm th}\right)_{ij} = \langle T_i^{(k)} T_j^{(k)} \rangle_k \, - \, \langle T_i^{(k)} \rangle_k \, \langle T_j^{(k)} \rangle_k, \tag{4.17}$$

in which the average is taken over PDF replicas. The two ways of propagating PDF uncertainties to the distribution of the EFT coefficients are equivalent assuming that PDF uncertainties are Gaussian and uncorrelated.

SIMUNET adopts the same optimisation settings as those set in the NNPDF analysis for the PDF-dependent part of the network. On the other hand it adjusts only those hyperparameters associated to the EFT-dependent layer. Within the joint SMEFT-PDF

Chapter 6.

fit, several of the fit settings such as the prior ranges for the EFT parameters and the learning rates are improved in an iterative way until convergence is achieved. In doing so, we also iterate the t_0 covariance matrix and the preprocessing exponents as is customary in the NNPDF procedure. In the fixed-PDF EFT fit, the user can decide both the ranges and the prior distributions to be used in the initial sampling of EFT coefficients as determined e.g. from a previous fit or from one-parameter scans.

4.3.2 New features

We now discuss some of the new features that have been implemented in SIMUnet, in comparison with [103], which are motivated by the needs of the SMEFT-PDF fits to LHC top quark data presented in this work. We consider in turn the following new features: the implementation of the quadratic contributions to the EFT cross-sections in the joint fits; fitting observables whose PDF dependence is negligible or non-existent; initialising the PDF weights of the neural network with the results of a previous fit; and finally, the improved initialisation of the EFT coefficients.

Quadratic EFT contributions. The version of SIMUnet used in [103] for the SMEFT-PDF fits of high-mass Drell-Yan data allowed the inclusion of quadratic contributions to the EFT cross-sections only under the approximation in which the cross-terms proportional to $c_i c_j$ with $i \neq j$ in Eq. (4.9) were neglected. In the current implementation, SIMUnet can instead account for the full quadratic contributions to the EFT cross-sections, including the non-diagonal cross-terms. This feature can be especially important for the interpretation of top quark measurements at the LHC, given that for many observables quadratic corrections, including cross-terms relating different operators, can be sizeable especially in the high-energy region.

The implementation consists of explicitly accounting for the cross terms, as parameters which depend on the Wilson coefficients and can be differentiated as a function of them during the training procedure.

PDF-independent observables. In the original version of SIMUnet, only physical observables with explicit dependence on both the PDFs (via the `FK`-tables interface) and the EFT coefficients could be included in a simultaneous fit. We have now extended the SIMUnet framework to describe observables that are independent of the PDF parameters $\boldsymbol{\theta}$, namely the weights and thresholds of the network depicted in Fig. 4.1 that output the SM partonic luminosity $\mathcal{L}^{(0)}$. For these PDF-independent observables, the SM predictions $T^{\rm SM}$ are evaluated separately and stored in theory tables which can be used to evaluate the SMEFT cross-sections after applying the rescaling of Eq. (4.9); hence, these observables only depend on the Wilson coefficients c_n.

In the current analysis the four-heavy cross-sections $\sigma_{\rm tot}(t\bar{t}b\bar{b})$ and $\sigma_{\rm tot}(t\bar{t}t\bar{t})$, the W-helicity measurements, and the associated top quark production cross-sections tZ, tW and $t\bar{t}\gamma$ are treated as PDF-independent observables, as for those cross-sections the PDF dependence can be neglected in comparison with other sources of theoretical and experimental uncertainty. The possibility to include PDF-independent observables makes SIMUNET a global SMEFT analysis tool on the same footing as SMEFIT [107, 43], FITMAKER [200], HEPFIT [281], EFTFITTER [282], and SFITTER [283] among others. This was explicitly demonstrated in App. C of Ref. [40], where it is shown that the results of a linear fixed-PDF SMEFT analysis performed with SIMUNET coincide with those obtained with SMEFIT [43] once the same experimental data and theory calculations are used (this benchmark is omitted here for brevity). Moreover, the new feature will allow us to include in future analyses any non-hadronic observables, such as electroweak precision observables (EWPO) [284].

Fixed-PDF weight initialisation. Within the current SIMUNET implementation, one can also choose to initialise the PDF-dependent weights of the network in Fig. 4.1 using the results of a previous Monte Carlo fit of PDFs, for example an existing SM-PDF analysis obtained with the NNPDF methodology. The weights of the latter are written to file and then read by SIMUNET for the network initialisation.

This feature has a two-fold application. First, instead of initialising at random the network weights in a simultaneous SMEFT-PDF fit, one can set them to the results of a previous SM-PDF fit, thus speeding up the convergence of the simultaneous fit, with the rationale that EFT corrections are expected to represent a perturbation of the SM predictions. Second, we can use this feature to compute EFT observables in the fixed-PDF case described above using the `FK`-table convolution with this previous PDF set as input, as opposed to having to rely on an independent calculation of the SM cross-section. Therefore, this PDF weight-initialisation feature helps realise SIMUNET both as a fixed-PDF EFT analysis framework, and to assess the stability of the joint SMEFT-PDF fits upon a different choice of initial state of the network in the minimisation.

Improved initialisation of the EFT coefficients. In the original implementation of SIMUNET it was only possible to initialise the EFT coefficients at specific values, selected beforehand by the user. In this work, we have developed more flexible initialisation schemes for the Wilson coefficients, in the sense that they can now be sampled from a prior probability distribution defined by the user; specifically, each Wilson coefficient c_i can be sampled from either a uniform $\mathcal{U}[a_i, b_i]$ or a normal $\mathcal{N}(\mu_i, \sigma_i)$ distribution. The ranges (a_i, b_i) of the uniform distribution $\mathcal{U}$ and the mean and standard deviation (μ_i, σ_i) of the Gaussian distribution $\mathcal{N}$ are now user-defined parameters, which can be assigned independently to each Wilson coefficients that enters the fit. This feature enhances the

effectiveness and flexibility of the minimisation procedure by starting from the regions of the parameter space with the best sensitivity to the corresponding Wilson coefficient.

Another option related to the improved initialisation of EFT coefficients is the possibility to adjust the overall normalisation of each coefficient by means of a user-defined scale factor. The motivation to introduce such a coefficient-dependent scale factor is to end up with (rescaled) EFT coefficients entering the fit which all have a similar expected range of variation. This feature is advantageous, since the resulting gradients entering the SGD algorithm will all be of the same order, and hence use a unique learning rate which is appropriate for the fit at hand.

4.4 Impact of the top quark Run II dataset on the SM-PDFs

Here we present the results of a global SM-PDF determination which accounts for the constraints of the most extensive top quark dataset considered to date in such analyses, described in Sect. 4.1. The fitting methodology adopted follows closely the settings of the NNPDF study [31]. This dataset includes not only the most up-to-date measurements of top quark pair production from Run II, but it also includes all available single top production cross-sections and distributions and for the first time new processes not considered in PDF studies before, such as the A_C asymmetry in $t\bar{t}$ production and the $t\bar{t}V$ and tV associated production (with $V = Z, W$).

We begin by summarising the methodological settings used for these fits in Sect. 4.4.1. Then in Sect. 4.4.2 we assess the impact of adding to a no-top baseline PDF fit various subsets of the top quark data considered in this study. In particular, we assess the impact of including updated Run II $t\bar{t}$ and single-top measurements in comparison with the subset used in the NNPDF analysis, see the second-to-last column of Tables 4.1– 4.7. Furthermore, we quantify for the first time the impact of associated vector boson and single-top (tV) as well as associated vector boson and top-quark pair production ($t\bar{t}V$) data in a PDF fit. Finally in Sect. 4.4.3 we combine these results and present a fit variant including all data described in Sect. 4.1, which is compared to both the NNPDF no-top baseline and to the original NNPDF set.

4.4.1 Fit settings

An overview of the SM-PDF fits that are discussed in this section is presented in Table 4.12. First of all, we produce a baseline fit, which we refer to as NNPDF no-top, which is based on the same dataset as NNPDF but with all top quark measurements excluded. Then we produce fit variants A to G, which quantify the impact of including in this baseline various

Fit ID	Datasets included in fit					
	No-top baseline, Sect. 4.1.1	**Incl. $t\bar{t}$, Table 4.1**	**Asymm., Table 4.2**	**Assoc. $t\bar{t}$, Table 4.4**	**Single-t, Table 4.7**	**Assoc. single$-t$, Table 4.8**
NNPDF4.0, no top (Baseline)	✓					
A (inclusive $t\bar{t}$)	✓	✓				
B (inclusive $t\bar{t}$ and charge asymmetry)	✓	✓	✓			
C (single top)	✓				✓	
D (all $t\bar{t}$ and single top)	✓	✓	✓		✓	
E (associated $t\bar{t}$)	✓			✓		
F (associated single top)	✓					✓
G (all associated top)	✓			✓		✓
H (all top data)	✓	✓	✓	✓	✓	✓

Table 4.12: Overview of the SM-PDF fits discussed in this section. The baseline fit, the no-top NNPDF4.0 fit, is based on the same dataset as NNPDF4.0 with all top quark measurements excluded. The fit variants A to G consider the impact of including in this baseline various subsets of top data, while in fit H the full set of top quark measurements described in Sect. 4.1 is added to the baseline.

subsets of top data (indicated by a check mark in the table). Finally, fit variant H is the main result of this section, namely the fit in which the full set of top quark measurements described in Sect. 4.1 is added to the no-top baseline.

In these fits, methodological settings such as network architecture, learning rates, and other hyperparameters are kept the same as in NNPDF, unless otherwise specified. One difference is the training fraction $f_{\rm tr}$ defining the training/validation split used for the cross-validation stopping criterion. In NNPDF, we used $f_{\rm tr} = 0.75$ for all datasets. Here instead we adopt $f_{\rm tr} = 0.75$ only for the no-top datasets and $f_{\rm tr} = 1.0$ instead for the top datasets. The rationale of this choice is to ensure that the fixed-PDF SMEFT analysis, where overfitting is not possible [201], exploits all the information contained in the top quark data considered in this study, and then for consistency to maintain the same settings in both the SM-PDF fits (this section) and in the joint SMEFT-PDF fits (to be discussed in Sect. 4.6). Nevertheless, we have verified that the resulting SM-PDF fits are statistically equivalent to the fits obtained by setting the training fraction to be 0.75 for all data, including for the top quark observables.

Fits A–G in Table 4.12 are composed of $N_{\rm rep} = 100$ Monte Carlo replicas after post-fit selection criteria, while the NNPDF no-top baseline fit and fit H are instead composed by $N_{\rm rep} = 1000$ replicas. As is customary, all fits presented in this section are iterated with respect to the t_0 PDF set and the pre-processing exponents.

4.4.2 Impact of individual top quark datasets

First we assess the impact of specific subsets of LHC top quark data when added to the NNPDF no-top baseline, fits A–G in Table 4.12. In the next section we discuss the

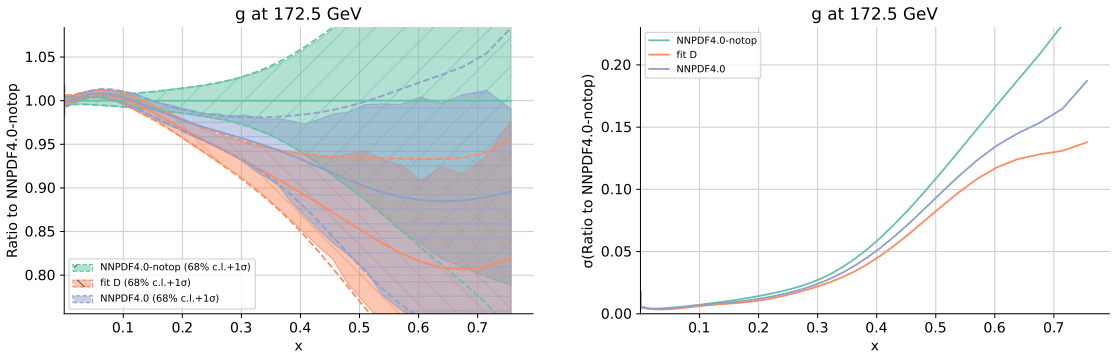


Figure 4.2: Left: comparison between the gluon PDF at $Q = m_t = 172.5$ GeV obtained in the NNPDF4.0 and NNPDF no-top fits against fit D in Table 4.12, which includes all top-quark pair (also the charge asymmetry A_C) and single-top quark production data considered in this analysis. Results are normalised to the central value of the NNPDF no-top set. Right: same comparison now for the PDF uncertainties (all normalised to the central value of the NNPDF no-top set).

outcome of fit H, which contains the full top quark dataset considered in this work.

Fig. 4.2 displays the comparison between the gluon PDF at $Q = m_t = 172.5$ GeV obtained in the NNPDF4.0 and NNPDF no-top fits against fit D, which includes all top-quark pair (also the charge asymmetry A_C) and all single-top quark production data considered in this analysis. Results are normalised to the central value of the NNPDF no-top fit, and in the right panel we show the corresponding PDF uncertainties, all normalised to the central value of the NNPDF no-top baseline. From Fig. 4.2 one finds that the main impact of the additional LHC Run II $t\bar{t}$ and single-top data included in fit D as compared to that already present in NNPDF4.0 is a further depletion of the large-x gluon PDF as compared to the NNPDF no-top baseline, together with a reduction of the PDF uncertainties in the same kinematic region. While fit D and NNPDF4.0 agree within uncertainties in the whole range of x, fit D and NNPDF no-top agree only at the 2σ level in the region $x \approx [0.2, 0.4]$. These findings imply that the effect on the SM-PDFs of the new Run II top data is consistent with, and strengthens, that of the data already part of NNPDF4.0, and suggests a possible tension between top quark data and other measurements in the global PDF sensitive to the large-x gluon, in particular inclusive jet and di-jet production. The reduction of the gluon PDF uncertainties from the new measurements can be as large as about 20% at $x \approx 0.4$. Differences are much reduced for the quark PDFs, and restricted to a 5% to 10% uncertainty reduction in the region around $x \sim 0.2$ with central values essentially unchanged.

To disentangle which of the processes considered dominates the observed effects on the gluon and the light quarks PDFs, Fig. 4.3 compares the relative PDF uncertainty on the gluon and on the d/u quark ratio in the NNPDF no-top baseline fit at $Q = m_t = 172.5$ GeV with the results from fits A, B, C, and D. As indicated in Table 4.12, these fit variants

include the following top datasets: inclusive $t\bar{t}$ (A), inclusive $t\bar{t}$ + A_C (B), single top (C), and their sum (D). The inclusion of the top charge asymmetry A_C data does not have any impact on the PDFs; indeed fits A and B are statistically equivalent. This is not surprising, given that in Eq. (4.1) the dependence on PDFs cancels out. Concerning the inclusion of single top data (fit C), it does not affect the gluon PDF but instead leads to a moderate reduction on the PDF uncertainties on the light quark PDFs in the intermediate-x region, $x \approx [0.01, 0.1]$, as shown in the right panel displaying the relative uncertainty reduction for the d/u ratio. This observation agrees with what was pointed out by a previous study [241], and the impact of LHC single-top measurements is more marked now as expected since the number of data points considered here is larger. We conclude that the inclusive $t\bar{t}$ measurements dominate the impact on the large-x gluon observed in Fig. 4.2, with single top data moderately helping to constrain the light quark PDFs in the intermediate-x region.

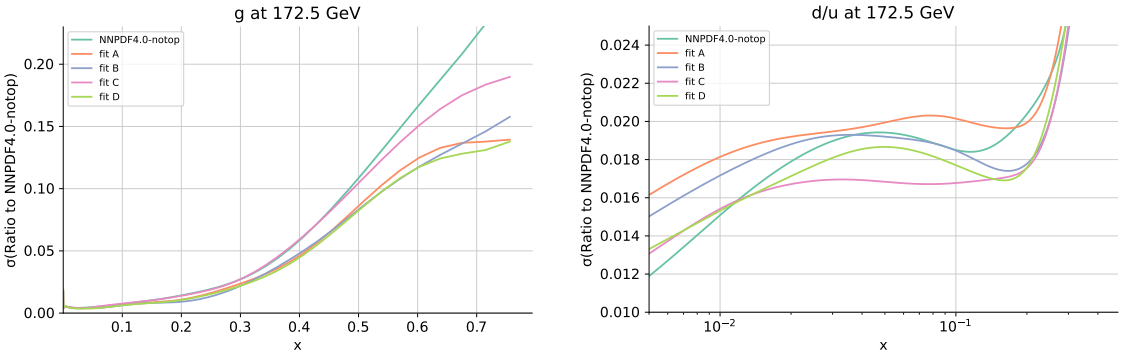


Figure 4.3: The ratio between the PDF 1σ uncertainty and the central value of the NNPDFnotop baseline in the case of the gluon (left panel) and in the case of the d/u ratio (right panel). The uncertainty of the baseline fit at $Q = m_t = 172.5$ GeV is compared with the uncertainty associated with fits A, B, C, and D in Table 4.12. These fit variants include the following top datasets: inclusive $t\bar{t}$ (A), inclusive $t\bar{t}$ + A_C (B), single top (C), and their sum (D).

We now consider the effect of the inclusion of processes that were not included before in any PDF fit, namely either $t\bar{t}$ or single-top production in association with a weak vector boson. Although current data exhibits large experimental and theoretical uncertainties, it is interesting to study whether they impact PDF fits at all, in view of their increased precision expected in future measurements; in particular, it is useful to know which parton flavours are most affected.

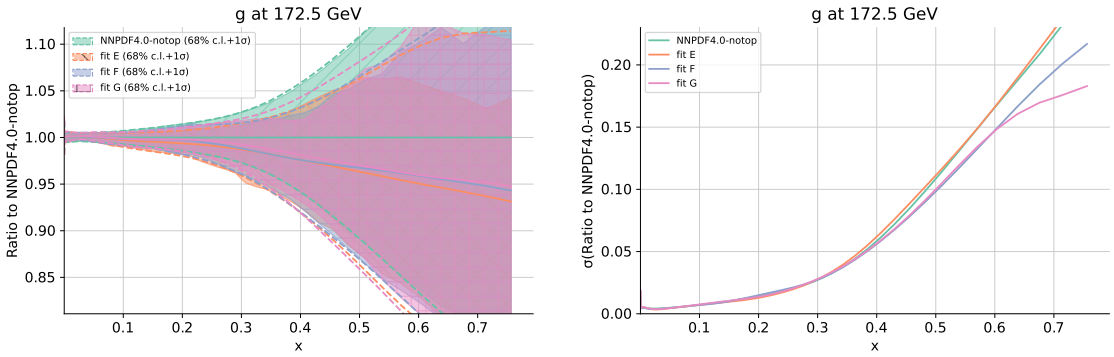


Figure 4.4: Same as Fig. 4.2 comparing the NNPDF no-top baseline fit with variants E, F, and G from Table 4.12. These variants include associated $t\bar{t}$ and vector boson production data (E), associated single top and vector boson production data (F) and their sum (G).

Fig. 4.4 displays the same comparison as in Fig. 4.2 now for the NNPDF no-top baseline and the variants E, F, and G from Table 4.12, which include the $t\bar{t}V$ (E) and tV (F) data as well as their sum (G). The pull of $t\bar{t}V$ is very small, while the pull of the tV measurements is in general small, but consistent with those of the inclusive $t\bar{t}$ measurements, namely preferring a depletion of the large-x gluon. This result indicates that $t\bar{t}V$ and tV data may be helpful in constraining PDFs once both future experimental data and theoretical predictions become more precise, although the corresponding inclusive measurements are still expected to provide the dominant constraints.

4.4.3 Combined effect of the full top quark dataset

The main result of this section is displayed in Fig. 4.5, which compares the NNPDF4.0 and the NNPDF no-top fits with variant H in Table 4.12, namely with the fit where the full set of top quark measurements considered in this analysis has been added to the no-top baseline. As in the case of Fig. 4.2, we show the large-x gluon normalised to the central value of NNPDF no-top and the associated 1σ PDF uncertainties (all normalised to the central value of the baseline). The results of fit H are similar to those of fit D, although slightly more marked. This is expected, since as shown above the associated production datasets $t\bar{t}V$ and tV carry little weight in the fit.

From Fig. 4.5 one observes how the gluon PDF of fit H deviates from the NNPDF no-top baseline markedly in the data region $x \in [0.1, 0.5]$. The shift in the gluon PDF can be up to the 2σ level, and in particular the two PDF uncertainty bands do not overlap in the region $x \in (0.2, 0.35)$. As before, we observe that the inclusion of the latest Run II top quark measurements enhances the effect of the top data already included in NNPDF4.0, by further depleting the gluon in the large-x region and by reducing its uncertainty by a factor up to 25%. Hence, one finds again that the new Run II top quark production measurements lead to a strong pull on the large-x gluon, qualitatively consistent but

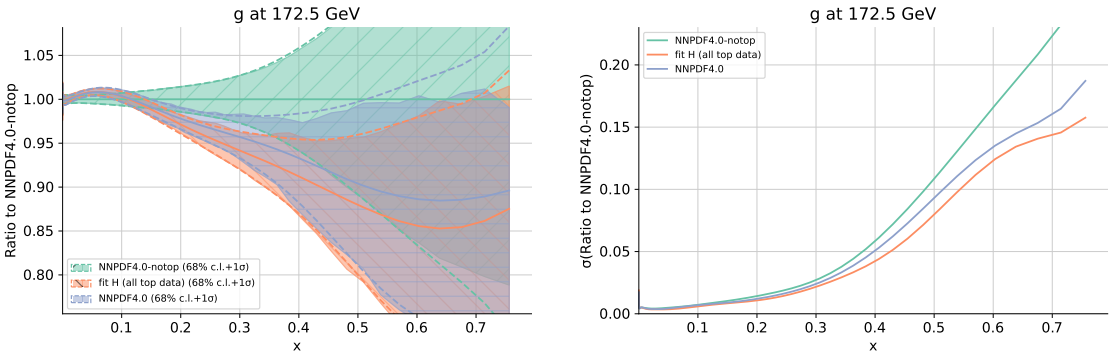


Figure 4.5: Same as Fig. 4.2 comparing NNPDF4.0 and NNPDF no-top with fit variant H in Table 4.12, which includes the full set of top quark measurements considered in this analysis.

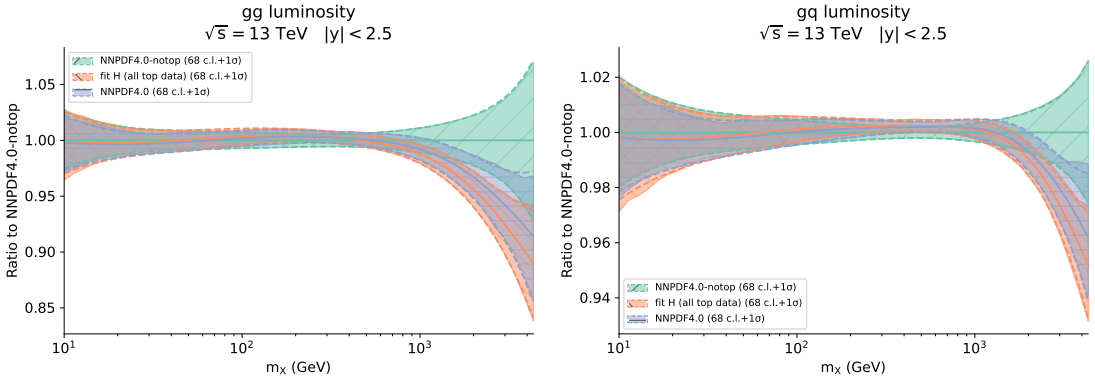


Figure 4.6: The gluon-gluon (left) and quark-gluon (right panel) partonic luminosities at $\sqrt{s} = 13$ TeV restricted to the central acceptance region with $|y| \leq 2.5$. We compare the NNPDF4.0 and NNPDF no-top fits with the predictions based on fit H, which includes the full top quark dataset considered here. Results are presented as the ratio to the central replica of the NNPDF no-top baseline fit.

stronger as compared with the pulls associated from the datasets already included in NNPDF4.0.

To assess the phenomenological impact of our analysis at the level of LHC processes, Fig. 4.6 compares the gluon-gluon and quark-gluon partonic luminosities at $\sqrt{s} = 13$ TeV (restricted to the central acceptance region with $|y| \leq 2.5$) between NNPDF4.0, NNPDF no-top, and fit H including the full top quark dataset considered here and Fig. 4.6 compares their uncertainties. Results are presented as the ratio to the no-top baseline fit. The qq and $q\bar{q}$ luminosities of fit H are essentially identical to those of the no-top baseline, as expected given the negligible changes in the quark PDFs observed in fit H, and hence are not discussed further.

From Figs. 4.6-4.7 one observes that both for the quark-gluon and gluon-gluon luminosity the impact of the LHC top quark data is concentrated on the region above $m_X \gtrsim 1$ TeV. As already reported for the case of the gluon PDF, also for the luminosities the

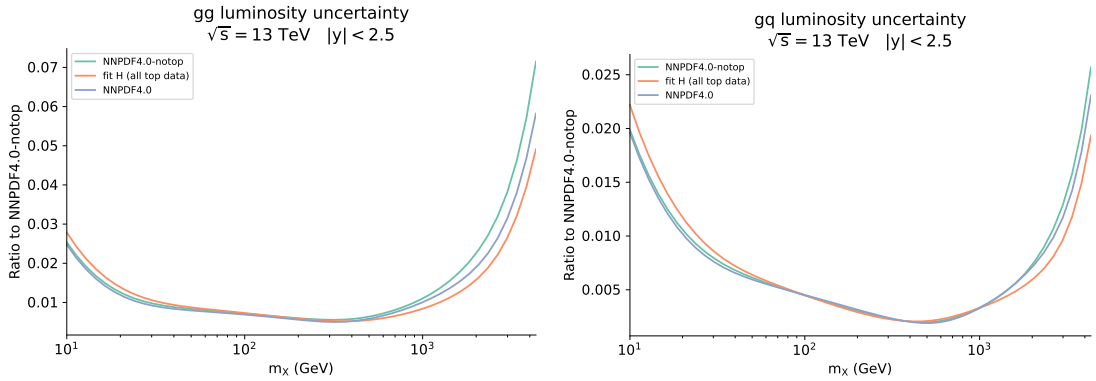


Figure 4.7: Same as Fig. 4.6, now for the relative luminosity uncertainties, all normalised to the NNPDF no-top baseline fit.

net effect of the new LHC Run II top quark data is to further reduce the luminosities for invariant masses in the TeV range, with a qualitatively similar but stronger pull as compared to that obtained in NNPDF4.0. While NNPDF4.0 and its no-top variant agree at the 1σ level in the full kinematical range considered, this is not true for fit H, whose error bands do not overlap with those of NNPDF no-top for invariant masses m_X between 2 and 4 TeV. On the other hand, NNPDF4.0 and fit H are fully consistent across the full m_X range, and hence we conclude that predictions for LHC observables based on NNPDF4.0 will not be significantly affected by the inclusion of the latest LHC top quark data considered in this work.

Finally, the fit quality of fit H is essentially stable, actually better relative to the NNPDF no-top baseline. The experimental χ^2 per data point on their respective datasets is 1.156 for the no-top baseline, whilst for fit H is reduced to 1.144. A complete summary of the χ^2 information for all of the fits in this section is given in App. D of Ref. [40] (the Appendix is omitted in this thesis for brevity). It is interesting to observe that all new top data included in fit H are already described well by using the NNPDF4.0 set, although clearly the χ^2 per data point improves (from 1.139 to 1.102) once all data are included in the fit. This confirms the overall consistency of the analysis.

4.5 Impact of the top quark Run II dataset on the SMEFT

We now quantify the impact of the LHC Run II legacy measurements, described in Sect. 4.1, on the top quark sector of the SMEFT. As compared to previous investigations of SMEFT operators modifying top-quark interactions [200, 201, 272, 285, 283, 286, 107, 287, 288, 289], the current analysis considers a wider dataset, in particular extended to various measurements based on the full LHC Run II luminosity. In the last column of

Tables 4.1–4.8 we indicated which of the datasets included here were considered for the first time within a SMEFT interpretation. Here we assess the constraints that the available LHC top quark measurements provide on the SMEFT parameter space, and in particular study the impact of the new measurements as compared to those used in [200, 201]. In this section we restrict ourselves to fixed-PDF EFT fits, where the input PDFs used in the calculation of the SM cross-sections are kept fixed. Subsequently, in Sect. 4.6, we generalise to the case in which PDFs are extracted simultaneously together with the EFT coefficients.

The structure of this section is as follows. We begin in Sect. 4.5.1 by describing the methodologies used to constrain the EFT parameter space both at linear and quadratic order in the EFT expansion. We also present results for the Fisher information matrix, which indicates which datasets provide constraints on which operators. In Sect. 4.5.2, we proceed to give the results of the fixed-PDF EFT fits at both linear and quadratic order, highlighting the impact of the new Run II top quark data by comparison with previous global SMEFT analyses. In Sect. 4.5.3, we assess the impact of replacing the CMS 13 TeV differential measurement of $t\bar{t}$ in the ℓ+jets channel, binned with respect to invariant top quark pair mass, by the corresponding double-differential measurement binned with respect to both invariant top quark pair mass and top quark pair rapidity. In the dataset selection performed in Sect 4.1.2 we rejected the double-differential distribution due to its poor χ^2-statistic in the SM, which could not be improved by a weighted fit of PDFs; in the present section, it is interesting to see whether the SMEFT can help account for the poor fit of this dataset. Finally, in Sect. 4.5.4 we evaluate the correlation between PDFs and EFT coefficients to identify the kinematic region and EFT operators which are potentially sensitive to the SMEFT-PDF interplay to be studied in Sect. 4.6.

4.5.1 Fit settings

Throughout this section, we will allow only the SMEFT coefficients to vary in the fit, keeping the PDFs fixed to the SM-PDFs baseline obtained in the NNPDF no-top fit discussed in Sect. 4.4; with this choice, one removes the overlap between the datasets entering the PDF fit and the EFT coefficients determination. Our analysis is sensitive to the $N = 25$ Wilson coefficients defined in Table 4.11, except at the linear level where the four-heavy coefficients c_{Qt}^8, c_{QQ}^1, c_{QQ}^8, c_{Qt}^1 and c_{tt}^1 (which are constrained only by $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ data) exhibit three flat directions [201]. In order to tackle this, we remove the five four-heavy coefficients from the linear fit.[3] Hence, in our linear fit we have $N = 20$

[3]In principle one could instead rotate to the principal component analysis (PCA) basis and constrain the two non-flat directions in the four-heavy subspace, but even so, the obtained constraints remain much looser in comparison with those obtained in the quadratic EFT fit [201].

In our fits, we also keep the $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ datasets after removing the five four-heavy coefficients. We have verified that including or excluding theses sets has no significant impact whatsoever on the remaining

independent coefficients constrained from the data, whereas in the quadratic fit, we fit all $N = 25$ independent coefficients.

The linear EFT fits presented in this section are performed with the SIMUnet methodology in the fixed-PDF option, as described in Sect. 5.1; we explicitly verified that the SIMUnet methodology reproduces the posterior distributions provided by SMEFiT (using either the NS (Nested Sampling) or MCfit options) for a common choice of inputs, as explicitly demonstrated in App. C of Ref. [40], but is omitted in this thesis for brevity. However, in the case of the quadratic EFT fits we are unable to use the SIMUnet methodology due to a failure of the Monte-Carlo sampling method utilised in the SIMUnet and SMEFiT codes; this is discussed in Sect. 4.7 and a dedicated study of the problem will be the subject of future work (with some preliminary discussion in Chapter 6). For this reason, quadratic EFT fits in this section are carried out with the public SMEFiT code using the NS mode [43]. To carry out these fits, the full dataset listed in Tables 4.1–4.8, together with the corresponding SM and EFT theory calculations described in Sect. 4.2, have been converted to the SMEFiT data format (this conversion was also already used for the benchmarking in App. C of Ref. [40]).

Fisher information. The sensitivity to the EFT operators of the various processes entering the fit can be evaluated by means of the Fisher information, F_{ij}, which quantifies the information carried by a given dataset on the EFT coefficients c_i [290]. In a linear EFT setting, the Fisher information is given by:

$$F_{ij} = L^{(i)T}(\text{cov}_{\text{exp}})^{-1}L^{(j)} \tag{4.18}$$

where the k-th entry, $L_k^{(i)}$, of the vector $L^{(i)}$ is the linear contribution multiplying c_i in the SMEFT theory prediction for the k-th data point, and cov_{exp} is the experimental covariance matrix. In particular, the Fisher information is an $N \times N$ matrix, where N is the number of EFT coefficients, and it depends on the dataset. An important property of the Fisher information is that it is related to the covariance matrix C_{ij} of the maximum likelihood estimators by the *Cramer-Rao bound*:

$$C_{ij} \geq (F^{-1})_{ij}, \tag{4.19}$$

indicating that larger values of F_{ij} will translate to tighter bounds on the EFT coefficients.

Before displaying the results of the fixed-PDF SMEFT analysis in Sect. 4.5.2, we use the Fisher information to assess the relative impact of each sector of top quark data on the EFT parameter space; this is done in the linear analysis, including $\mathcal{O}(1/\Lambda^2)$ SMEFT corrections. In the quadratic case, once $\mathcal{O}(1/\Lambda^4)$ SMEFT corrections are included, the

coefficients.

dependence of F_{ij} on the Wilson coefficients makes interpretation more difficult. Writing $F_{ij}(D)$ for the Fisher information matrix evaluated on the dataset D, we define the relative constraining power of the dataset D via:

$$\text{relative constraining power of } D \text{ on operator } c_i = F_{ii}(D) \Big/ \sum_{\text{sectors } D'} F_{ii}(D'). \qquad (4.20)$$

Since $F_{ii}(D)$ corresponds to the constraining power of the dataset D in a one-parameter fit of the Wilson coefficient c_i, this definition only quantifies how much a dataset impacts one-parameter fits of single Wilson coefficients in turn; however, this will give a general qualitative picture of some of the expected behaviour in the global fit too. We display the results of evaluating the relative constraining power of each top quark data sector on each of the parameters in Fig. 4.8, quoting the results in percent (%).

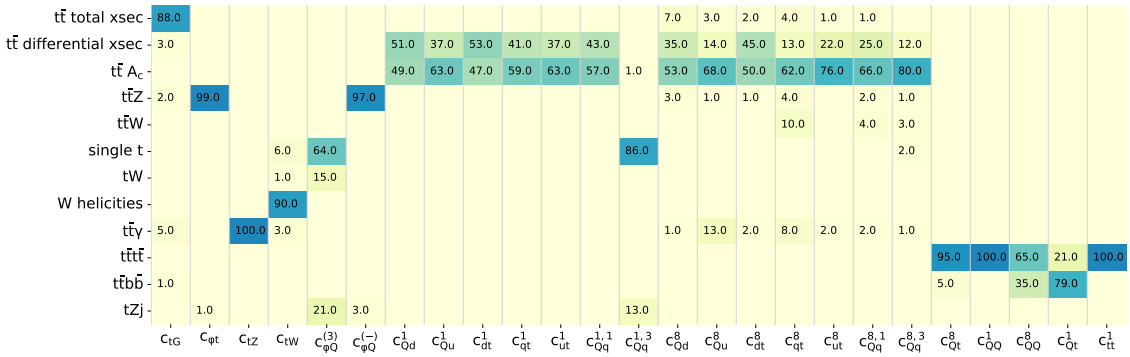


Figure 4.8: Relative constraining power (in %) on each of the operators for each of the processes entering the fit, as defined in Eq. (4.20).

As expected, $t\bar{t}$ total cross sections constitute the dominant source of constraints on the coefficient c_{tG}. Each of the four-fermion operators receive important constraints from differential $t\bar{t}$ distributions and charge asymmetry measurements. Note that this impact is magnified when we go beyond individual fits, in which case measurements of charge asymmetries are helpful in breaking flat directions amongst the four-fermion operators [283, 200]. The coefficient $c_{Qq}^{1,3}$ is the exception as it is instead expected to be well-constrained by single top production. We note that the measurements of W helicities are helpful in constraining the coefficient c_{tW}, while $t\bar{t}Z$ measurements provide the dominant source of constraints on $c_{\varphi Q}^{(-)}$. We observe that the neutral top coupling c_{tZ} is entirely constrained by $t\bar{t}\gamma$, and that the effects of $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ are mostly restricted to the 4-heavy operators c_{Qt}^8, c_{QQ}^1, c_{QQ}^8, c_{Qt}^1 and c_{tt}^1.

4.5.2 Fixed-PDF EFT fit results

In this section, we present the results of the linear and quadratic fixed-PDF fits with the settings described in Sect 4.5.1.

Fit quality. We begin by discussing the fit quality of the global SMEFT determination, quantifying the change in the data-theory agreement relative to the SM in both the linear and quadratic SMEFT scenarios. Table 4.13 provides the values of the χ^2 per data point in the SM and in the case of the SMEFT at both linear and quadratic order in the EFT expansion for each of the processes entering the fit. Here, in order to ease the comparison of our results to those of SMEFiT and FITMAKER, we quote the χ^2 per data point computed by using the covariance matrix defined in Eq. (4.16), which includes both the experimental uncertainty and the PDF uncertainty. The corresponding values obtained by using the experimental χ^2 definition of Eq. (4.3), along with a fine-grained fit quality description are given in App. D of Ref. [40].

We observe that in many sectors, the linear EFT fit improves the fit quality compared to the SM; notably, the $\chi^2_{\rm exp+th}$ per data point for inclusive $t\bar{t}$ is vastly improved from 1.71 to 1.11. When quadratic corrections are also considered, the fit quality is usually poorer compared to the linear fit. For example, in inclusive $t\bar{t}$ the $\chi^2_{\rm exp+th}$ per data point deteriorates from 1.11 to 1.69. This is not unexpected, however, since the flexibility of the quadratic fit is limited by the fact that for sufficiently large values of Wilson coefficients the EFT can only make positive corrections.[4]

It is also useful to calculate the goodness of fit, quantified by the the χ^2 per degree of freedom, $\chi^2/n_{\rm dof} = \chi^2/(n_{\rm dat} - n_{\rm param})$, which additionally accounts for the complexity of the models we are using in each fit. In our case, we find $\chi^2_{\rm exp+th}/n_{\rm dof} = 1.25$ in the SM and $\chi^2_{\rm exp+th}/n_{\rm dof} = 0.95$ and 1.33 in the linear and quadratic EFT scenarios respectively. We see that while the EFT at quadratic order does not provide a better fit than the SM, neglecting quadratic EFT corrections leads to a significant improvement in the overall fit quality.

Constraints on the EFT parameter space. Next, we present the constraints on the EFT parameter space. In Fig. 4.9, we display the 95% CL constraints on the 20 Wilson coefficients entering the linear fit. Two sets of constraints are shown; in green, we give the intervals obtained from a fit to the 175 data points introduced in Sect. 4.1, whilst in orange, we give the intervals obtained from a fit to the older top quark dataset used in the global analysis of Ref. [201], obtained from a fit of 150 data points. This comparison allows us to quantify the information gained from the latest Run II datasets, relative to

[4]This also has methodological implications. Large quadratic corrections can negatively impact the Monte-Carlo sampling method used by SIMUNET, as discussed in Sect. 4.7.

Process	$n_{\rm dat}$	$\chi^2_{\rm exp+th}$ [SM]	$\chi^2_{\rm exp+th}$ [SMEFT $\mathcal{O}(\Lambda^{-2})$]	$\chi^2_{\rm exp+th}$ [SMEFT $\mathcal{O}(\Lambda^{-4})$]
$t\bar{t}$	86	1.71	1.11	1.69
$t\bar{t}$ AC	18	0.58	0.50	0.60
W helicities	4	0.71	0.45	0.47
$t\bar{t}Z$	12	1.19	1.17	0.94
$t\bar{t}W$	4	1.71	0.46	1.66
$t\bar{t}\gamma$	2	0.47	0.03	0.59
$t\bar{t}t\bar{t}$ & $t\bar{t}b\bar{b}$	8	1.32	1.06	0.49
single top	30	0.504	0.33	0.37
tW	6	1.00	0.82	0.82
tZ	5	0.45	0.30	0.31
Total	**175**	**1.24**	**0.84**	**1.14**

Table 4.13: The values of the χ^2 per data point for the fixed-PDF EFT fits presented in this section, both for individual groups of processes, and for the total dataset. Here the χ^2 is actually the $\chi^2_{\rm exp+th}$ defined by using the theory covariance matrix defined in Eq. (4.16). In each case we indicate the number of data points, the $\chi^2_{\rm exp+th}$ obtained using the baseline SM calculations, and the results of both the linear and quadratic EFT fits.

those available to previous analyses. The same comparison, this time at quadratic order in the EFT expansion, is shown in Fig. 4.10 (note that in this plot we display constraints on all 25 coefficients, including the 4-heavy coefficients c_{Qt}^8, c_{QQ}^1, c_{QQ}^8, c_{Qt}^1 and c_{tt}^1).

We first note that Figures 4.9 and 4.10 both demonstrate good agreement between the fits using old and new datasets, and consistency between the new fit SMEFT bounds and the SM. At the linear level, the most noticeable improvement concerns c_{tG}; its 95% C.L. bounds decrease from $[-0.13, 0.41]$ to $[-0.18, 0.17]$, thanks to the increased amount of information in the input dataset, coming in particular from $t\bar{t}$ data. This results in both a tightening of the constraints by about 35% and a shift in the best-fit point. For many of the other coefficients, the bounds are either stable (e.g. c_{tZ}), or exhibit a shift in central value but no significant tightening (e.g. $c_{\varphi t}$, undergoes a shift of -14.3, but a decrease in the size of the constraints 95% C.L. by only 1%, and $c_{\varphi Q}^{(-)}$ undergoes a shift of -7.33, but its bounds only tighten by 2%). Finally, we note that some coefficients instead exhibit a broadening of constraints with the new dataset relative to the old dataset (for example, some of the four-fermion operators). The increase in the size of the constraints could point to some inconsistency within the new inclusive $t\bar{t}$ dataset; however, given that

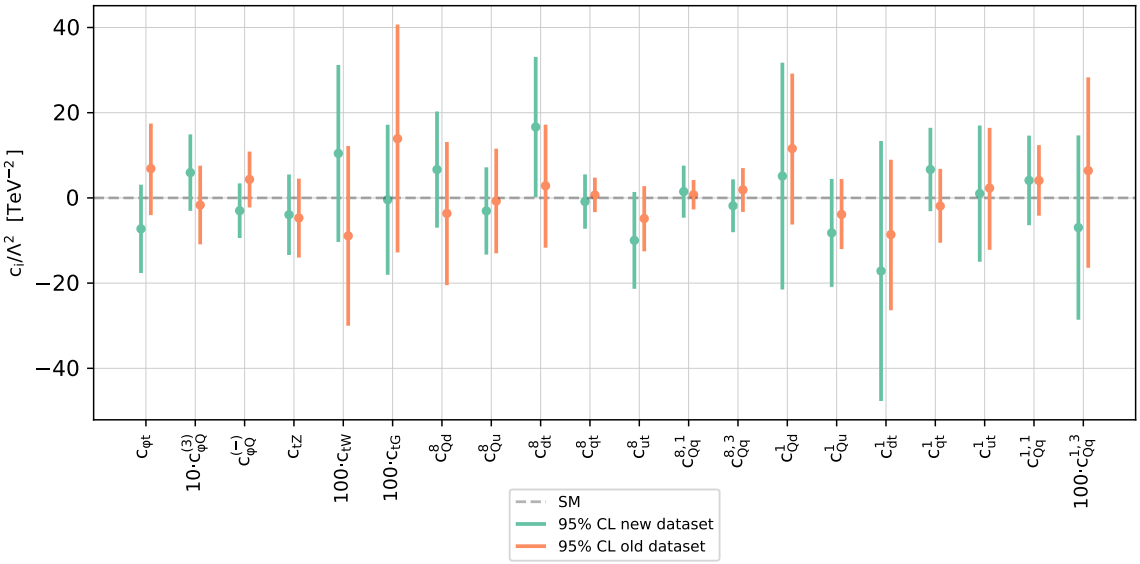


Figure 4.9: The 95% CL intervals on the EFT coefficients entering the linear fit, evaluated with SIMUNET on the dataset considered in this work, and evaluated with SMEFIT on the top quark dataset entering the analysis of [201] (note that at the linear EFT level, the results obtained with SIMUNET coincide with those provided by SMEFITfor the same dataset, as demonstrated in App. C of Ref. [40]). Note also that the constraints on selected coefficients are rescaled by the factors shown, for display purposes.

the bounds are very large anyway, at the edges of the intervals we are likely to approach a region where the EFT is no longer valid in both cases, hence no definite conclusions may be drawn.

At the quadratic order in the EFT expansion, however, the impact of the latest Run II dataset is clear; we see a marked improvement in many of the SMEFT constraints. As shown in Fig. 4.10, the bounds on all 14 of the four-fermion operators become noticeably smaller as a result of the increase in precision in the $t\bar{t}$ sector. The constraint on c_{tZ} is improved by the inclusion of measurements of the $t\bar{t}\gamma$ total cross sections, resulting in a tightening of 24%. The addition of the p_T^γ spectrum [240] would yield an even stronger constraint, as seen in [200]. We will make use of this observable in future work when unfolded measurements are made available. Contrary to the linear fit, where we singled out $c_{\varphi t}$ and $c_{\varphi Q}^{(-)}$ as examples of coefficients which shift, but whose bounds are not improved, the constraints on $c_{\varphi t}$, $c_{\varphi Q}^{(-)}$ markedly tighten in the quadratic fit in the presence of new data; in particular, the size of the bounds on $c_{\varphi t}$, $c_{\varphi Q}^{(-)}$ decrease by 35% and 28%, respectively. On the other hand, despite the addition of new $t\bar{t}t\bar{t}$, $t\bar{t}b\bar{b}$ datasets, we find limited sensitivity to the five four-heavy coefficients c_{Qt}^8, c_{QQ}^1, c_{QQ}^8, c_{Qt}^1 and c_{tt}^1. In fact, with the new data we see a broadening of the bounds. As with the linear fit, this could point to either an inconsistency in the $t\bar{t}t\bar{t}$, $t\bar{t}b\bar{b}$ data, or simply to the ambiguity associated to the EFT validity in that particular region of the parameter space.

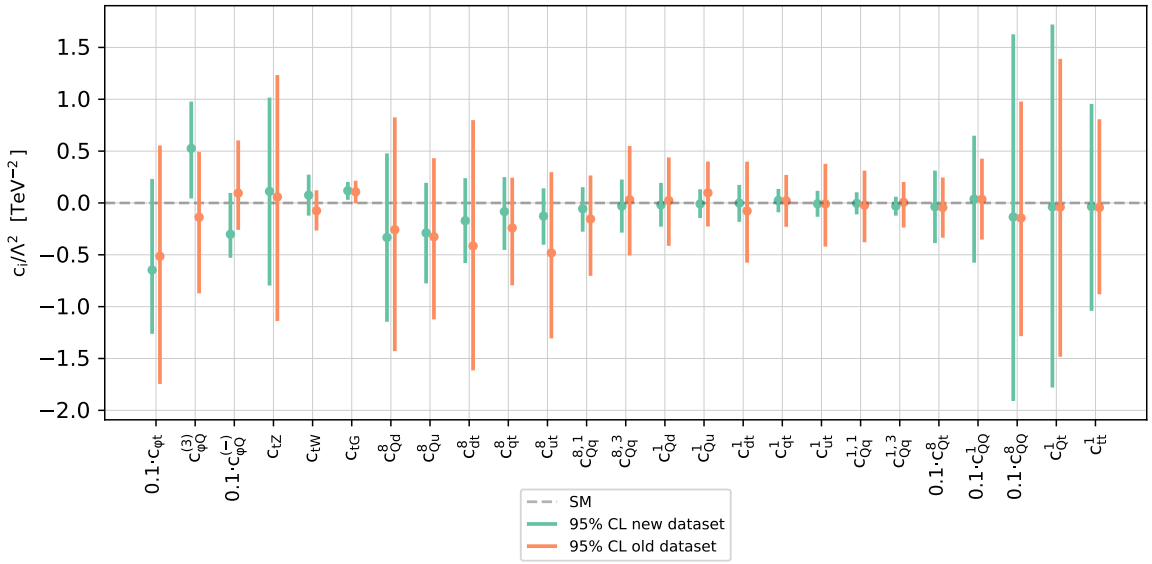


Figure 4.10: The 95% CL intervals on the EFT coefficients entering the quadratic fit, evaluated with SMEFiT. We compare the results based on the full top quark dataset with the corresponding results obtained from the subset of top quark measurements entering the analysis of [201]. As in Fig. 4.9, the constraints on selected coefficients are rescaled by the factors shown, for display purposes.

Correlations. Figure 4.11 shows the correlations between Wilson coefficients evaluated in this analysis both at the linear and the quadratic order in the EFT expansion, shown on the left and right panels respectively. In the linear fit, we first note a number of large correlations amongst the octet four-fermion operators which enter the $t\bar{t}$ production together. The singlet four-fermion operators are similarly correlated among themselves, although their correlations are comparatively suppressed. The coefficients $c_{\varphi t}$ and $c_{\varphi Q}^{(-)}$ exhibit a large positive correlation due to their entering into $t\bar{t}Z$ production together, while $c_{Qq}^{3,1}$ and $c_{\varphi Q}^{(3)}$ have positive correlations through their contribution to single top production. Further non-zero correlations are found, for example amongst the pairs $c_{Qq}^{8,3}$ & $c_{\varphi Q}^{(3)}$, c_{Qd}^{8} & c_{tZ} and c_{Qu}^{8} & c_{tZ}.

At quadratic order, however, we observe that many of these correlations are suppressed, as a result of the fact that the inclusion of quadratic corrections lifts many of the degeneracies in the fit. We observe that the pairs $c_{\varphi t}$, $c_{\varphi Q}^{(-)}$ and $c_{Qq}^{1,3}$, $c_{\varphi Q}^{(3)}$ remain correlated though. The 4-heavy operators are also included in this quadratic fit, and we find large anti-correlations between c_{QQ}^{1} and c_{QQ}^{8}, indicating that they are poorly distinguished in the $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ processes. Finally, note that we obtain subtle non-zero correlations between the octet and singlet four-fermion operators constructed from the same fields, for example between c_{Qd}^{8} and c_{Qd}^{1}. This is a result of the fact that $t\bar{t}$ measurements provide the dominant source of constraints on these coefficients and are very sensitive to quadratic corrections, and at this order the contribution from these operators differs only by a

numerical factor.

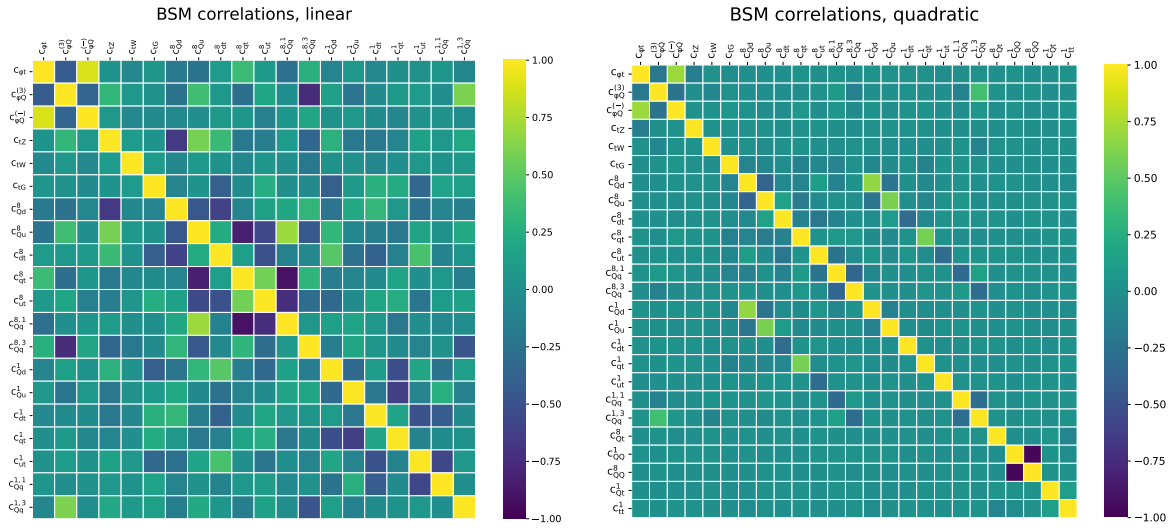


Figure 4.11: The values of the correlation coefficients evaluated between all pairs of Wilson coefficients entering the EFT fit at the linear (left) and quadratic (right) order. As explained in the next, the number of fitted DoFs is different in each case.

4.5.3 Study of the CMS 1D vs 2D distribution

In the dataset selection discussed in Sect. 4.1.2, the double-differential distribution measurements performed by CMS at $\sqrt{s} = 13$ TeV in the ℓ+jets channel [217] – binned with respect to the top quark pair invariant mass $m_{t\bar{t}}$ and the top quark pair rapidity $y_{t\bar{t}}$ – is found to be poorly described by the SM theory, with an experimental χ^2 that is 22σ away from the median of the χ^2 distribution of a perfectly consistent dataset made of $n_{\rm dat} = 34$ points. Even by increasing the weight of this dataset in a weighted fit (see Sect. 4.1.2 for a more detailed explanation), the $\chi^2_{\rm exp}$ per data point improves only moderately to 4.56 and the χ^2-statistic of the other datasets deteriorates significantly, hinting to both an internal incompatibility of the CMS 2D distribution and to an incompatibility with the rest of the data. For this reason, the dataset was excluded from our analysis, and replaced by the single-differential distribution in $m_{t\bar{t}}$ – presented in the same publication [217]. The latter is well described by the SM theoretical predictions.

In this section we present a dedicated analysis to assess whether the SMEFT corrections can improve the theoretical description of this dataset. In particular, we compare a fixed-PDF fit including the 13 TeV CMS double-differential $(m_{t\bar{t}}, y_{t\bar{t}})$ distribution (CMS 2D) to the default one including the 13 TeV single-differential $m_{t\bar{t}}$ distribution (CMS 1D).

First of all, it is interesting to notice that the inclusion of quadratic SMEFT corrections in the fit does not significantly improve the quality of the fit of the CMS 2D distribution, with $\chi^2_{\rm exp+th}/n_{\rm dat}$ decreasing from 2.80 (in the SM) to $\chi^2_{\rm exp+th}/n_{\rm dat} = 2.57$ including SMEFT

Process	$n_{\rm dat}$	$\chi^2_{\rm exp+th}$ [SM]		$\chi^2_{\rm exp+th}$ [SMEFT $\mathcal{O}(\Lambda^{-2})$]		$\chi^2_{\rm exp+th}$ [SMEFT $\mathcal{O}(\Lambda^{-4})$]	
		CMS 1D	CMS 2D	CMS 1D	CMS 2D	CMS 1D	CMS 2D
$t\bar{t}$	86	1.71	2.07	1.11	1.18	1.69	1.87
$t\bar{t}$ AC	18	0.58	0.58	0.50	0.47	0.60	0.60
W helicities	4	0.71	0.71	0.45	0.45	0.48	0.46
$t\bar{t}Z$	12	1.19	1.19	1.17	1.07	0.94	0.95
$t\bar{t}W$	4	1.71	1.71	0.46	0.46	1.66	1.82
$t\bar{t}\gamma$	2	0.47	0.47	0.03	0.03	0.58	0.18
$t\bar{t}t\bar{t}$ & $t\bar{t}b\bar{b}$	8	1.32	1.32	1.06	1.28	0.49	0.49
single top	30	0.50	0.50	0.33	0.34	0.37	0.35
tW	6	1.00	01.00	0.82	0.86	0.82	0.84
tZ	5	0.45	0.45	0.30	0.30	0.31	0.30
Total	**175**	**1.24**	**1.48**	**0.84**	**0.91**	**1.14**	**1.29**

Table 4.14: Same as Table 4.13, now comparing the values of the $\chi^2_{\rm exp+th}$ per data point for the fixed-PDF EFT fits presented in this section, the default one including the 13 TeV single-differential $m_{t\bar{t}}$ distribution (CMS 1D) and the one including the 13 TeV CMS double-differential $(m_{t\bar{t}}, y_{t\bar{t}})$ distribution (CMS 2D), both for individual groups of processes, and for the total dataset.

$\mathcal{O}(\Lambda^{-4})$ corrections. On the other hand, if SMEFT linear $\mathcal{O}(\Lambda^{-2})$ corrections are included, the fit quality of the CMS 2D distribution improves substantially with $\chi^2_{\rm exp+th}/n_{\rm dat} = 1.22$.

In order to assess the effect of the inclusion of the CMS 2D distribution on the other top sector data, in Table 4.14 we compare the fit quality of the two (fixed-PDF) SMEFT fits, CMS 1D and CMS 2D, both for individual groups of processes, and for the total dataset. We observe that the fit quality deteriorates both in the SM and in the quadratic SMEFT fit once the CMS 2D distribution is fitted. The deterioration of the fit quality is mostly driven by a deterioration of the fit quality of the $t\bar{t}$ and $t\bar{t}W$ sectors. However at the level of linear SMEFT fit, the quality of the fit deteriorates only moderately and it is mostly driven by the slight deterioration in the inclusive $t\bar{t}$ sector and in the $t\bar{t}t\bar{t}\,\&\,t\bar{t}b\bar{b}$ one.

At the level of the fit of the Wilson coefficients, the quadratic SMEFT fits yields similar 95% C.L. bounds on the EFT coefficients. This is somewhat expected, given that the fit quality of the CMS 2D data does not improve once quadratic SMEFT corrections are included, and the fit quality of the other datasets remains pretty much the same. On the other hand the bounds obtained from a SMEFT linear fit change more significantly depending on whether the CMS 1D or the CMS 2D distribution is used in the fit. In

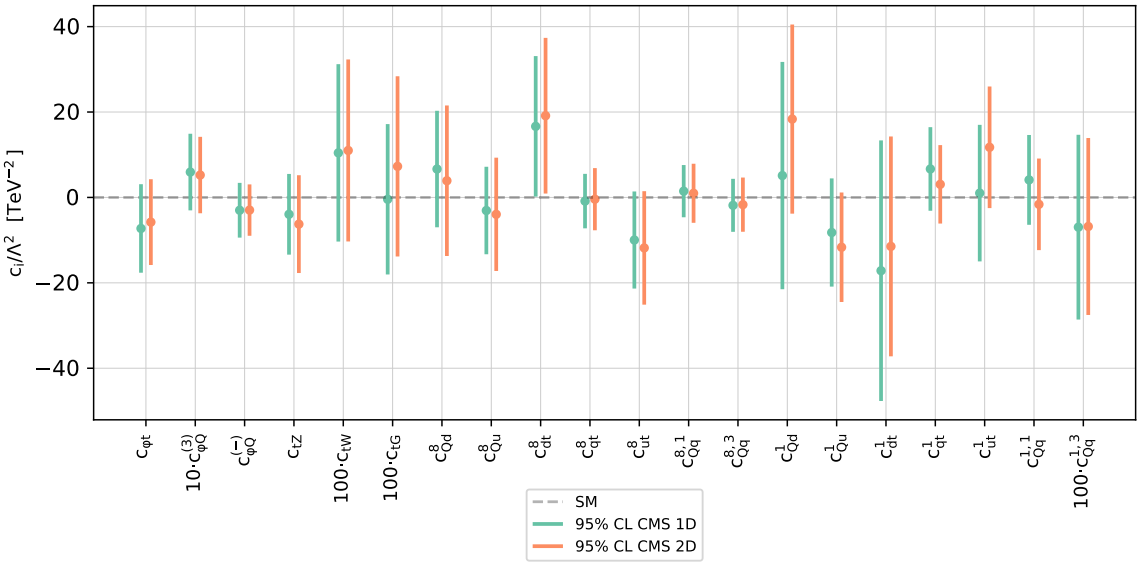


Figure 4.12: The 95% CL intervals on the EFT coefficients entering the liner fits, evaluated with SIMUNET. We compare the results based on the default full top quark dataset (CMS 1D) with the corresponding results obtained from the same set of top quark measurements in which the single-differential $m_{t\bar{t}}$ 13 TeV CMS distribution is substituted by the double differential $(m_{t\bar{t}}, y_{t\bar{t}})$ 13 TeV CMS distribution (CMS 2D).

particular the central value of the c_{tG} 95% C.L. bounds is shifted upwards by 1σ, as well as the bounds on c_{Qd}^1 and c_{ut}^1, while the bounds on c_{Qu}^1 and c_{qt}^1 shift downwards, as expected from the EFT coefficient correlations displayed in Fig. 4.11.

As a result of our analysis, we observe that the bounds on the EFT in the linear case do depend on the input dataset quite significantly and careful consideration has to be made to the overall dataset compatibility and to the outcome of the fit once quadratic corrections are included. The current analysis in particular shows that the incompatibility of the 13 TeV CMS double differential distribution cannot be entirely attributed to new physics effects parametrised by the SMEFT expansion, rather there are internal experimental or theoretical incompatibilities that affect the results.

To conclude, it is interesting to observe that, if one tries to include the CMS double differential distribution dataset instead of the single differential dataset in a simultaneous fit of PDFs and SMEFT at the linear level, the improvement in the fit quality that is obtained in a SMEFT linear fit is reverted. Indeed, the SMEFT PDF resulting from the simultaneous fit including the CMS double differential distribution (instead of the CMS single differential one) are basically unchanged with respect to the ones that will be presented in Sect. 6, and the bounds on the SMEFT coefficients are quite similar. However, the fit quality of the CMS double differential distribution is rather poor, with $\chi^2_{\rm exp}/n_{\rm dat} = 5.69$ (only marginally decreasing from 6.43 in the PDF-only fit). This means that, in a simultaneous fit, the degrees of freedom of the PDFs tend to contrast the

direction that the SMEFT coefficients would take in the absence of PDFs because the latter would require a change in the PDFs that is disfavoured by the other datasets in the global analysis. All results can be found on a public web page that includes the additional material that we do not show in this chapter.[5]

4.5.4 Correlations between PDFs and EFT coefficients

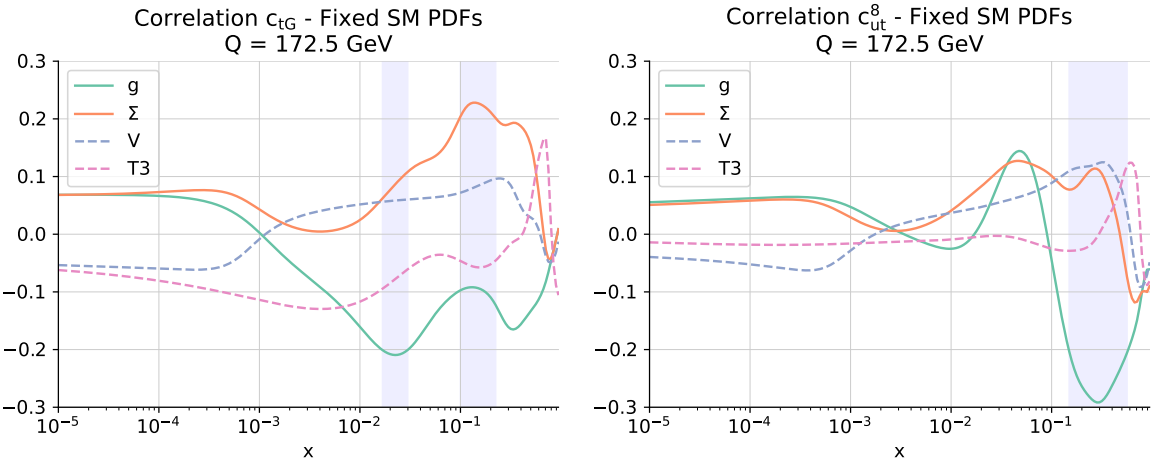


Figure 4.13: The value of the correlation coefficient ρ between the PDFs and selected EFT coefficients as a function of x and $Q = 172.5$ GeV. We show the results for the gluon, the total singlet Σ, total valence V, and non-singlet triplet T_3 PDFs. We provide results for representative EFT coefficients, namely c_{tG} and $c_{ut}^{(8)}$.

We conclude this section by discussing the correlations observed between the PDFs and Wilson coefficients. The PDF-EFT correlation coefficient for a Wilson coefficient c and a PDF $f(x, Q)$ at a given x and Q^2 is defined as

$$\rho\left(c, f(x,Q^2)\right) = \frac{\left\langle c^{(k)} f^{(k)}(x,Q^2)\right\rangle_k - \left\langle c^{(k)}\right\rangle_k \left\langle f^{(k)}(x,Q^2)\right\rangle_k}{\sqrt{\langle (c^{(k)})^2\rangle_k - \langle c^{(k)}\rangle_k^2}\sqrt{\left\langle \left(f^{(k)}(x,Q^2)\right)^2\right\rangle_k - \langle f^{(k)}(x,Q^2)\rangle_k^2}}\,, \tag{4.21}$$

where $c^{(k))}$ is the best-fit value of the Wilson coefficient for the k-th replica and $f^{(k)}$ is the k-th PDF replica computed at a given x and Q, and $\langle\cdot\rangle_k$ represents the average over all replicas. We will compute the correlation between a SM PDF and the Wilson coefficients, both of which have been separately determined from the total dataset including all new top quark data. By doing so we hope to shed light on which Wilson coefficients, and which PDF flavours and kinematical regions, are strongly impacted by the top quark data and therefore exhibit a potential for interplay in a simultaneous EFT-PDF determination. The EFT corrections will be restricted to linear order in the EFT expansion.

[5] `https://www.pbsp.org.uk/topproject/`

Fig. 4.13 displays a selection of the largest correlations. We observe that the gluon PDF in the large-x region is significantly correlated with the Wilson coefficients c_{tG}, $c_{ut}^{(8)}$. On the other hand, relatively large correlations are observed between c_{tG} and the the total singlet Σ, while the total valence V, and non-singlet triplet T_3 PDFs show no relevant correlations with the selected coefficients. This is not surprising, given the impact of top quark pair production total cross sections and differential distributions in constraining these PDFs and Wilson coefficients. Whilst these correlations are computed from a determination of the SMEFT in which the PDFs are fixed to SM PDFs, the emergence of non-zero correlations provides an indication of the potential for interplay between the PDFs and the SMEFT coefficients; this interplay will be investigated in a simultaneous determination in the following section.

4.6 SMEFT-PDFs from top quark data

In this section we present the main results of this chapter, namely the simultaneous determination of the proton PDFs and the SMEFT Wilson coefficients from the LHC Run II top quark data described in Sect. 4.1, following the SIMUNET methodology summarised in Sect. 5.1. This determination of the SMEFT-PDFs from top quark data is carried out at the linear, $\mathcal{O}(1/\Lambda^2)$, level in the EFT expansion. We do not perform simultaneous fits at the quadratic level due to shortcomings of the Monte Carlo replica method, on which SIMUNET is based; this is discussed in detail in Sect. 4.7, and is also the subject of Chapter 6.

PDFs from a joint SMEFT-PDF fit. We begin by discussing the PDFs obtained through a joint fit of PDFs and Wilson coefficients from the complete LHC top quark dataset considered in this work. Simultaneously extracting the PDFs and the EFT coefficients from top quark data has a marked impact on the former, as compared to a SM-PDF baseline, but we shall see has much less impact on the latter, as compared to the results of the corresponding fixed-PDF EFT analyses. Fig. 4.14 displays a comparison between the gluon and quark singlet PDFs, as well as of their relative 1σ PDF uncertainties, for the no-top baseline, the SM-PDFs of fit H in Table 4.12 which include the full top quark dataset, and their SMEFT-PDF counterparts based on the same dataset. PDFs are compared in the large-x region for $Q = m_t = 172.5$ GeV. In the left panel they are normalised to the central value of the no-top fit.

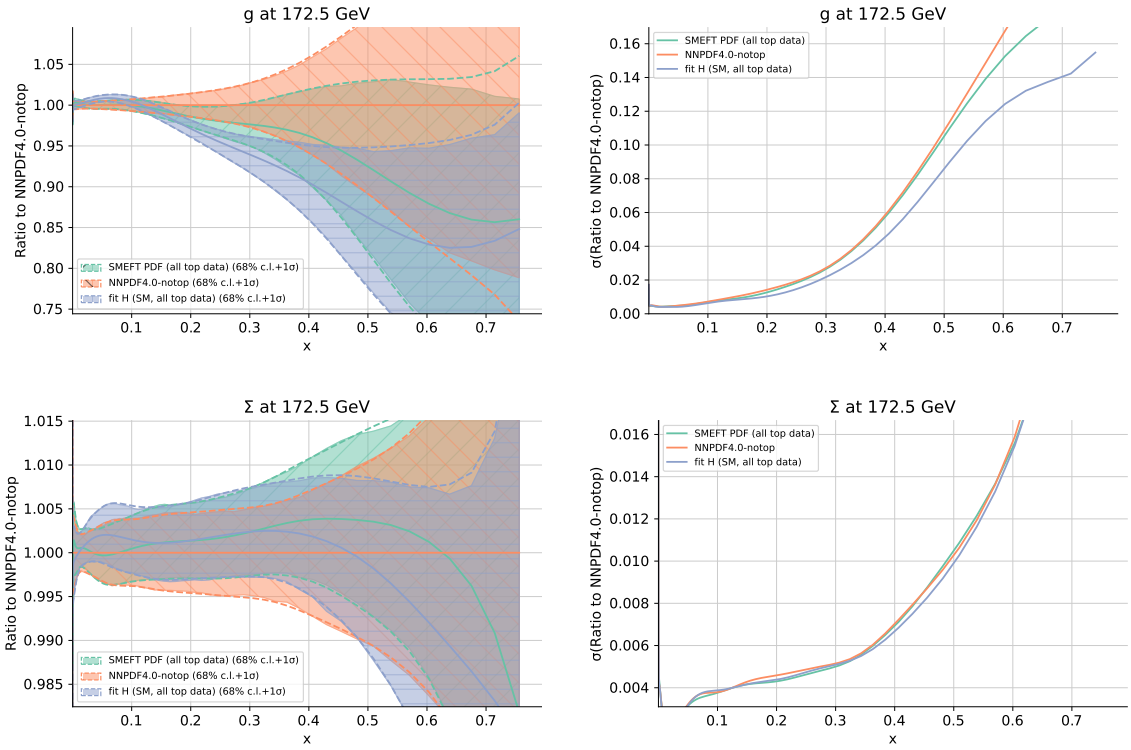


Figure 4.14: Left: a comparison between the gluon (upper panel) and quark singlet (lower panel) PDFs evaluated at $Q = m_t = 172.5$ GeV in the large-x region. We display the no-top baseline, the SM-PDFs of fit H in Table 4.12 which include the full top quark dataset, and their SMEFT-PDF counterparts. The results are normalised to the central value of the no-top fit. Right: the same comparison, but now for the relative 1σ PDF uncertainties.

While differences are negligible for the quark singlet PDF, both in terms of central values and uncertainties, they are more marked for the gluon PDF. Two main effects are observed therein. First, the central value of the SMEFT-PDF gluon moves upwards as compared to the SM-PDF fit based on the same dataset, ending up halfway between the latter and the no-top fit. Second, uncertainties increase for the SMEFT-PDF determination as compared to the SM-PDFs extracted from the same data, becoming close to the uncertainties of the no-top fit except for $x \gtrsim 0.5$. In both cases, differences are restricted to the large-x region with $x \gtrsim 0.1$, where the impact of the dominant top quark pair production measurements is the largest.

The results of Fig. 4.14 for the gluon PDF therefore indicate that within a simultaneous extraction of the PDFs and the EFT coefficients, the impact of the top quark data on the PDFs is diluted, with the constraints it provides partially 'reabsorbed' by the Wilson coefficients. This said, there remains a pull of the top quark data as compared to the no-top baseline fit which is qualitatively consistent with the pull obtained in a SM-PDF determination based on the same dataset, albeit of reduced magnitude. Interestingly, as we show below, while the SMEFT-PDF gluon is significantly modified in the joint fit as

compared to a SM-PDF reference, much smaller differences are observed at the level of the bounds on the EFT parameters themselves.

Fig. 4.15 displays the same comparison as in Fig. 4.14 now for the case of the gluon-gluon and quark-gluon partonic luminosities at $\sqrt{s} = 13$ TeV, as a function of the final-state invariant mass m_X. Consistently with the results obtained at the PDF level, one finds that the three luminosities are almost identical for $m_X \leq 1$ TeV, and at higher invariant masses the SMEFT-PDF predictions are bracketed between the no-top fit from above and the SM-PDF which includes all top data (fit H in Table 4.12) from below. Hence, the net effect of simultaneously fitting the PDFs and the EFT coefficients is a dilution of constraints provided by the top quark data on the former, which translates into larger PDF uncertainties (which end up being rather similar to those of no-top) and an increase in the large-m_X luminosity, e.g. of 5% in the gg case for $m_X \simeq 3$ TeV, as compared to the SM-PDF luminosity.

This dilution arises because of an improved description of the top quark data included in the fit, especially in the high $m_{t\bar{t}}$ bins. In the SM-PDF case this can only be obtained by suppressing the large-x gluon, while in a SMEFT-PDF analysis this can also be achieved by shifting the EFT coefficients from their SM values. In other words, as compared to the no-top baseline, the gluon PDF experiences a suppression at large-x of up to 10% when fitting the top quark data, and this pull is reduced by approximately a factor two in the joint SMEFT-PDF determination due to the coherent effect of the linear EFT cross-sections.

It is worth mentioning that, since we include the SMEFT corrections by applying BSM factors computed bin-by-bin by taking ratios of the SMEFT contributions and the SM cross sections with a specific PDF set, a change of the PDFs can in principle translate into a change of these factors. Our current methodology relies therefore on performing this check a posteriori and in the case of modified BSM factors, the fit is reiterated with the new ones. However, in the present case, we find that the BSM factors are substantially unaffected and we do not reiterate the fit. The parton luminosities are indeed very similar to the ones of the PDF set `NNPDF40_nnlo_as_01180` used for the original calculation, with deviations of $2-3\%$ at most. At the current experimental sensitivity, changes of $\mathcal{O}(1\%)$ in the EFT corrections can be safely ignored as they do not impact in a significant way the fits.

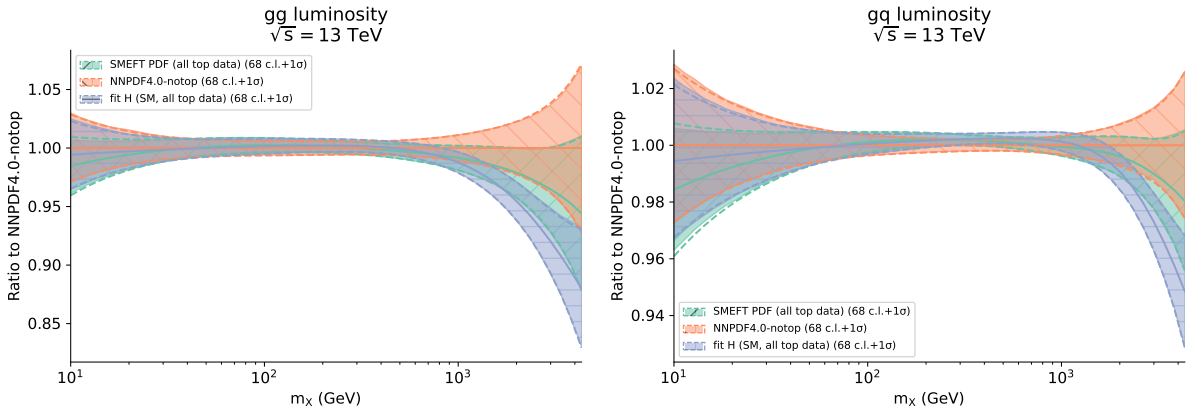


Figure 4.15: The gluon-gluon (left panel) and quark-gluon (right panel) partonic luminosities at $\sqrt{s} = 13$ TeV as a function of the final-state invariant mass m_X. We compare the no-top baseline fit with its SM-PDF counterpart including all top quark data considered (fit H in Table 4.12) as well as with the SMEFT-PDF determination. Results are presented as the ratio to the central value of the no-top baseline.

EFT coefficients from a joint SMEFT-PDF fit. As opposed to the marked effect of the SMEFT-PDF interplay found for the large-x gluon, its impact is more moderate at the level of the bounds on the EFT coefficients, and is restricted to mild shifts in the central values and a slight broadening of the uncertainties. This is illustrated by Fig. 4.16, showing the posterior distributions for the Wilson coefficient c_{tG} associated to the chromomagnetic operator in the joint SMEFT-PDF determination, compared with the corresponding results from the fixed-PDF EFT analysis whose settings are described in Sect. 4.5. The comparison is presented both for the fits which consider only top-quark pair production data and those based on the whole top quark dataset considered in this work. The leading effect of the chromomagnetic operator $\mathcal{O}_{tG}$ is to modify the total rates of $t\bar{t}$ production without altering the shape of the differential distributions, and hence it plays an important role in a simultaneous SMEFT-PDF determination based on top quark data.

For fits based only on inclusive $t\bar{t}$ data, as shown in the left panel of Fig. 4.16, the two posterior distributions are similar; the distribution based on the SMEFT-PDF analysis is slightly broader, approximately 10% so, as compared to the fixed-PDF EFT fit to the same measurements. This slight broadening is washed out in the fit to the full top quark dataset, as shown in the right panel of Fig. 4.16. In both cases, the determination of c_{tG} is consistent with the SM at a 95% CL, and the best-fit values of the coefficient are the same in the SMEFT-PDF and fixed-PDF EFT analyses. Hence, in the specific case of the chromomagnetic operator, the interplay between PDFs and EFT fits is rather moderate and restricted to a broadening of at most 10% in the 95% CL bounds. Similar comparisons have been carried out for other EFT coefficients as well as in the context of fits to a subset of the data and/or to a subset of the coefficients. We find that in general the impact of the SMEFT-PDF interplay translates to a broadening of the uncertainties in the EFT

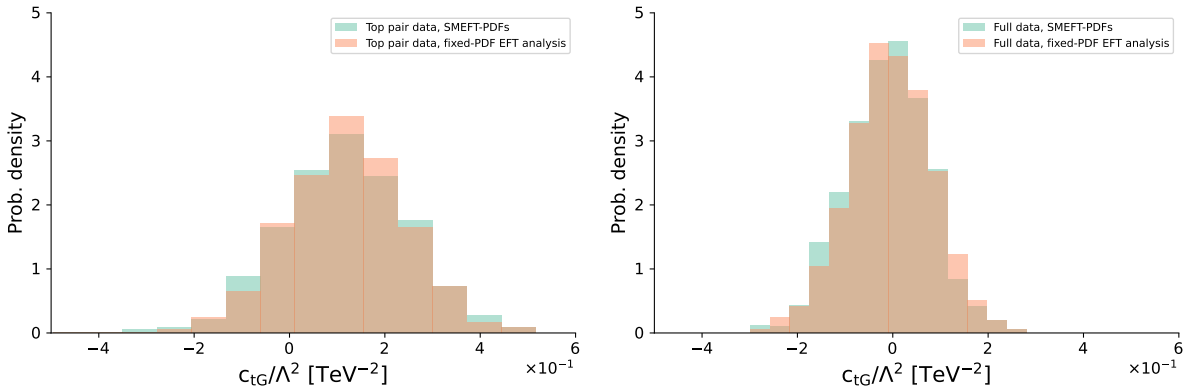


Figure 4.16: Posterior distributions for the Wilson coefficient c_{tG} associated to the chromomagnetic operator in the joint SMEFT-PDF determination, compared with the corresponding results from the fixed-PDF EFT analysis whose settings are described in Sect. 4.5. We show results based on only top-quark pair production data (left) and in the whole top quark dataset considered in this work (right panel).

coefficients, which at most reaches 30%, and alongside which the best-fit values remain stable.[6]

All in all, within the global fit based on the best available theory predictions, results for the EFT coefficients turn out to be very similar in the fixed-PDF EFT and SMEFT-PDF fits. This indicates that, provided a broad enough dataset and the most up-to-date theory calculations are used, the PDF dependence on the cross-sections entering an EFT interpretation of the LHC data is currently subdominant and can be neglected (this is not the case for the PDFs, see Fig. 4.15). Nevertheless, this statement applies only to the dataset currently available, and it is likely that the SMEFT-PDF interplay will become more significant in the future once HL-LHC measurements become available [291, 292], as demonstrated in the case of high-mass Drell-Yan [102] in Chapter 3.

The moderate impact of the SMEFT-PDF interplay on the Wilson coefficients for the full top quark dataset considered in this work is summarised in Fig. 4.17, which compares the 95% CL intervals of the 20 fitted Wilson coefficients relevant for the linear EFT fit obtained from the outcome of the joint SMEFT-PDF determination and the fixed-PDF EFT analysis. The latter is based on SM and EFT calculations performed with no-top as input; see also the description of the settings in Sect. 4.5. The dashed horizontal line indicates the SM prediction, and some coefficients are multiplied by the indicated prefactor to facilitate the visualisation of the results. Fig. 4.17 demonstrates that, other than slight broadenings and shifts in the central values, the results of the two analyses coincide.

Correlations. Fig. 4.18 displays the correlation coefficients [277] between the SMEFT-PDFs and the Wilson coefficients evaluated at $Q = 172.5$ GeV as a function of x. Each

[6]All results obtained with various subsets of the data and of the coefficients can be found at `https://www.pbsp.org.uk/research/topproject`

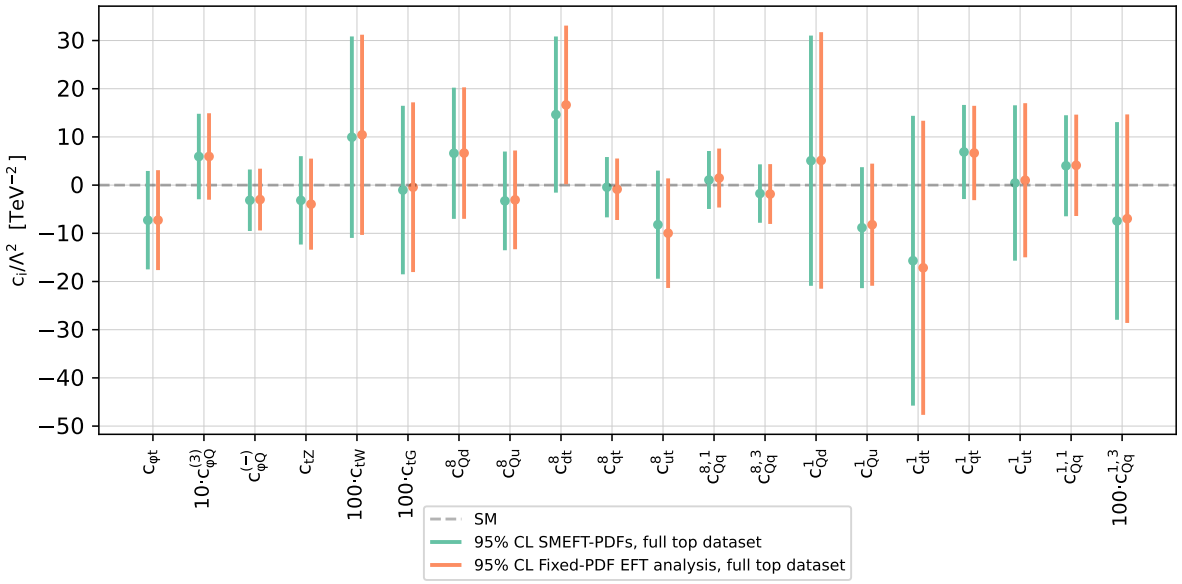


Figure 4.17: Comparison of the 95% CL intervals on the 20 Wilson coefficients considered in this chapter (in the linear EFT case) between the outcome of the joint SMEFT-PDF determination and that of the fixed-PDF EFT analysis. The latter is based on SM and EFT calculations performed with no-top as input, see also Sect. 4.5. In both cases, results are based on the full top quark dataset being considered and EFT cross-sections are evaluated up to linear, $\mathcal{O}\left(\Lambda^{-2}\right)$, corrections. The dashed horizontal line indicates the SM prediction, $c_k = 0$. Note that some coefficients are multiplied by the indicated prefactor to facilitate the visualisation of the results.

panel displays the correlations of the coefficient c_k with the gluon and the total singlet Σ, total valence V, and non-singlet triplet T_3 PDFs, and we consider four representative EFT coefficients, namely c_{td}^8, c_{tu}^8, c_{tG}, and c_{tW}. The largest correlations within the EFT coefficients considered in this work are associated to the gluon PDF and four-fermion operators such as c_{td}^8 and c_{tu}^8 in the large-x region, peaking at $x \simeq 0.3$. Correlations for other values of x and for the quark PDFs are negligible for all operators entering the analysis. We note that future data with an enhanced coverage of the high-$m_{t\bar{t}}$ region in top quark pair-production might alter this picture, given that for $m_{t\bar{t}} \gtrsim 3$ TeV the $q\bar{q}$ luminosity starts to become more relevant and eventually dominates over the gg contribution.

Residuals. Finally, Fig. 4.19 displays a similar comparison as in Fig. 4.17 now at the level of the 68% CL fit residuals defined as

$$R_n = \frac{c_n^*}{\sigma_n}\,, \tag{4.22}$$

where c_n^* and σ_n are the median value and the standard deviation of the Wilson coefficient c_n respectively, with $n = 1, \ldots, N$, where N is the number of operators. The outcome of the joint SMEFT-PDF determination is compared with that of a fixed-PDF EFT analysis

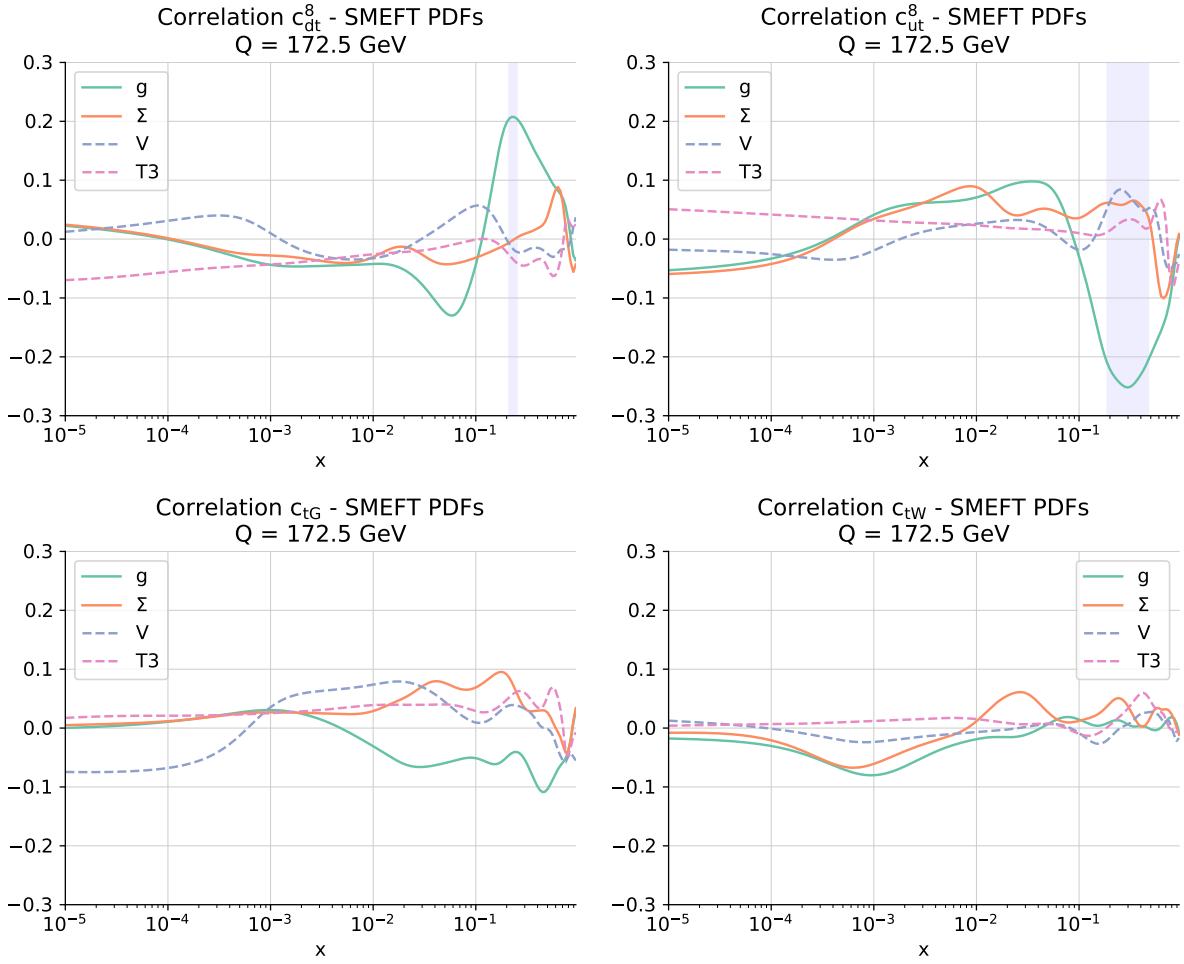


Figure 4.18: The correlation coefficient $\rho(f_i, c_k)$ between the SMEFT-PDFs f_i and the Wilson coefficients c_k evaluated at $Q = 172.5$ GeV as a function of x. Each panel displays the correlations of the coefficient c_k with the gluon and the total singlet Σ, total valence V, and non-singlet triplet T_3 PDFs. We provide results for representative EFT coefficients, namely c^8_{td}, c^8_{tu}, c_{tG}, and c_{tW}. The largest correlations within the EFT coefficients considered in this work are associated to four-fermion operators such as c^8_{td} and c^8_{tu}.

where we use as input for the theory calculations the SMEFT-PDFs obtained in the joint fit, rather than the no-top set. That is, in both cases the information provided by the top quark data on the PDFs and Wilson coefficients is accounted for, but in one case the cross-correlations are neglected whereas they are accounted for in the other. The residuals are similar in the two cases; they are slightly bigger (in absolute value) in the fixed-PDF case in which the correlations between the SMEFT-PDFs and the EFT coefficients are ignored. This analysis further emphasises that, for the currently available top quark data, the cross-talk between PDFs and EFT degrees of freedom does not significantly modify the posterior distributions in the space spanned by the Wilson coefficients.

In summary, on the one hand we find that from the point of view of a PDF determination,

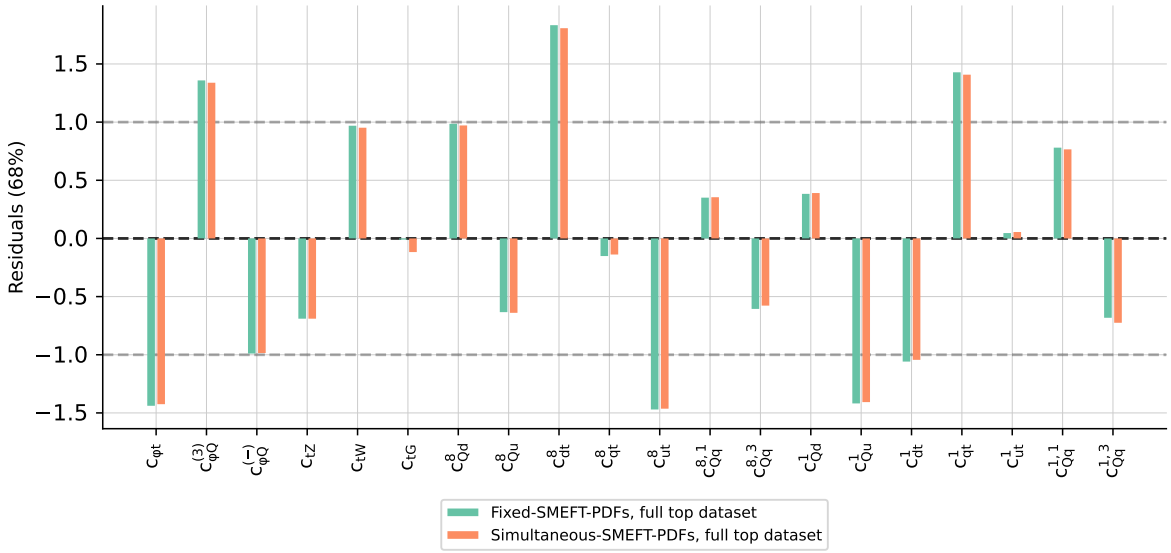


Figure 4.19: The 68% CL residuals, Eq. (4.22), for the same Wilson coefficients displayed in Fig. 4.17, comparing the outcome of the joint SMEFT-PDF determination with that of a fixed-PDF EFT analysis. In the latter, we use as input for the theory calculations the SMEFT-PDFs obtained in the joint fit rather than the no-top set used in Sect. 4.5. The horizontal dashed lines indicate the $\pm 1\sigma$ regions.

SM-PDFs and SMEFT-PDFs extracted from top quark data differ by an amount comparable to their respective uncertainties in the case of the large-x gluon. On the other hand, at the level of Wilson coefficients the results are unchanged irrespective of the PDF set used as input for the theory calculations; that is, bounds based on the no-top fit or the SMEFT-PDFs are almost the same. Hence, while EFT interpretations of top quark data can safely ignore the PDF dependence, at least for the settings adopted in this work, a global PDF fit could be significantly distorted if BSM physics were to be present in the large-energy top quark distributions.

4.7 Pitfalls of the Monte-Carlo replica method for quadratic EFT fits

The analysis presented in this chapter was originally prepared with the intention of performing a fully simultaneous SMEFT-PDF fit using NLO QCD theory, including *quadratic*, $\mathcal{O}\left(\Lambda^{-4}\right)$, contributions from the SMEFT. To this end, the capabilities of the SIMUNET framework were extended, as discussed in Sect. 4.3.2, and all necessary SMEFT K-factors needed for the quadratic predictions were produced. However, whilst benchmarking our code, we noticed significant disagreement between quadratic SMEFT-only fits produced using the SIMUNET methodology and the SMEFIT Nested Sampling

option;[7] on the other hand, we note perfect agreement between our SIMUnet quadratic SMEFT-only fits and the SMEFiT MCfit option.

This disagreement can be traced back to a deficiency in the Monte-Carlo sampling method used to propagate experimental error to the SMEFT coefficients, which currently prevents us from applying the the SIMUnet framework to joint SMEFT-PDF fits with quadratic EFT calculations. In this section, we describe our current understanding of these limitations within the context of a toy model, and give a more realistic example from this work; further work on the topic is deferred to a future publication, with some preliminary discussion given in Chapter 6.

4.7.1 A toy model for quadratic EFT fits

In the following subsection, we consider a toy scenario involving a single data point d and only one Wilson coefficient c (we ignore all PDF-dependence). We suppose that our observed experimental data point d is a random variable drawn from a normal distribution centred on the underlying quadratic theory prediction, and with experimental variance σ^2:

$$d \sim N(t(c), \sigma^2). \tag{4.23}$$

We assume that the theory prediction is quadratic in the SMEFT Wilson coefficient c, taking the form:

$$t(c) = t^{\mathrm{SM}} + ct^{\mathrm{lin}} + c^2 t^{\mathrm{quad}}\,, \tag{4.24}$$

where we set $\Lambda = 1$ TeV for convenience. Recall that $t^{\mathrm{quad}} > 0$, since it corresponds to a squared amplitude. Given the observed data d, we would like to construct interval estimates for the parameter c (usually *confidence intervals* in a frequentist setting or *credible intervals* in a Bayesian setting). Here, we shall describe the analytical construction of two interval estimates: first, using the Bayesian method, and second, using the Monte-Carlo replica method.

Bayesian method. In the Bayesian approach, c is treated as a random variable with its own distribution. By Bayes' theorem, we can write the probability distribution of c, given the observed data d, up to a proportionality constant (given by $1/\mathbb{P}(d)$, where $\mathbb{P}(d)$ is called the *Bayes' evidence*) as:

$$\mathbb{P}(c|d) \propto \mathbb{P}(d|c)\mathbb{P}(c), \tag{4.25}$$

[7] This disagreement is not present at the linear level; in this case, SIMUnet using the fixed-PDF option and SMEFiTusing either the Nested Sampling or MCfit options perfectly coincide. See App. C of Ref. [40] for more details.

where $\mathbb{P}(c|d)$ is called the *posterior distribution* of c, given the observed data d, and $\mathbb{P}(c)$ is called the *prior distribution* of c - this is our initial 'best guess' of the distribution of c before the observation of the data takes place. This distribution is often taken to be uniform in SMEFT fits; we shall assume this here.

Given a value of c, the distribution $\mathbb{P}(d|c)$ of the data d is assumed to be Gaussian, as specified in Eq. (4.23). In particular, we can deduce that the posterior distribution of c obeys the following proportionality relation:[8]

$$\mathbb{P}(c|d) \propto \exp\left(-\frac{1}{2\sigma^2}\left(d - t(c)\right)^2\right). \tag{4.26}$$

This posterior distribution can be used to place interval estimates on the parameter c. One way of doing this is to construct *highest density intervals*. These are computed as follows. For a $100\alpha\%$ credible interval, we determine the constant $p(\alpha)$ satisfying:

$$\int\limits_{\{c:\ \mathbb{P}(c|d)>p(\alpha)\}} \mathbb{P}(c|d)\ dc = \alpha. \tag{4.27}$$

An interval estimate for c is then given by $\{c : \mathbb{P}(c|d) > p(\alpha)\}$. In order to obtain such intervals then, we must construct the posterior $\mathbb{P}(c|d)$; efficient sampling from the posterior is guaranteed by methods such as Nested Sampling [293, 294].

The Monte-Carlo replica method. This method takes a different approach in order to produce a posterior distribution for the parameter c. Given the observed central data value d, one samples repeatedly from the normal distribution $N(d, \sigma^2)$ to generate a collection of *pseudodata replicas*, which we shall denote as $d^{(1)}, ..., d^{(N_{\text{rep}})}$, where N_{rep} is the total number of replicas. Given a pseudodata replica $d^{(i)}$, one obtains a corresponding best-fit value of the Wilson coefficient parameter $c^{(i)}$ by minimising the χ^2 of the theory to the pseudodata:

$$c^{(i)} = \arg\min_c \chi^2(c, d^{(i)}) = \arg\min_c \left(\frac{(d^{(i)} - t(c))^2}{\sigma^2}\right). \tag{4.28}$$

In this toy scenario, we can determine an analytical formula for $c^{(i)}$:

$$c^{(i)} = \begin{cases} -\dfrac{t^{\text{lin}}}{2t^{\text{quad}}}, & \text{if } d^{(i)} \leq \left(t^{\text{SM}} - (t^{\text{lin}})^2/4t^{\text{quad}}\right)\ ; \\ \dfrac{-t^{\text{lin}} \pm \sqrt{(t^{\text{lin}})^2 - 4t^{\text{quad}}(t^{\text{SM}} - d^{(i)})}}{2t^{\text{quad}}}, & \text{if } d^{(i)} \geq \left(t^{\text{SM}} - (t^{\text{lin}})^2/4t^{\text{quad}}\right)\ . \end{cases} \tag{4.29}$$

The first case arises when the χ^2 to the pseudodata $d^{(i)}$ has a single minimum, whilst the second case arises when the χ^2 has two minima. The two cases are split based on the

[8] Technically, truncated according to the end-points of the uniform prior.

value of

$$t_{\min} = t^{\rm SM} - (t^{\rm lin})^2/4t^{\rm quad}\,, \tag{4.30}$$

which is the minimum value of the quadratic theory prediction $t(c) = t^{\rm SM} + ct^{\rm lin} + c^2 t^{\rm quad}$. Note that for data replicas such that $d^{(i)} \leq t_{\min}$, the best-fit value $c^{(i)}$ becomes independent of $d^{(i)}$ and depends only on the ratio between linear and quadratic EFT cross-sections.

Now, given that $d^{(i)}$ is a random variable drawn from the normal distribution $N(d, \sigma^2)$, one can infer the corresponding distribution of the random variable $c^{(i)}$, which is a function of the pseudodata $d^{(i)}$. For a real random variable X with associated probability density $P_X(x)$, a function $f : \mathbb{R} \to \mathbb{R}$ of the random variable has the distribution:

$$P_{f(X)}(y) = \int_{-\infty}^{\infty} dx\ P_X(x)\delta(y - f(x)). \tag{4.31}$$

In our case, $c^{(i)}$ is a multi-valued function of $d^{(i)}$ given the two square roots, but the formula is easily generalised to this case. Recalling that the pseudodata replicas are generated according to a Gaussian distribution around the central measurement d with variance σ^2, we find that the probability density function for the Wilson coefficient replica $c^{(i)}$ is given (up to a proportionality constant) by:

$$\begin{aligned} P_{c^{(i)}}(c) \propto & \int_{-\infty}^{t_{\min}} dx\ \delta\left(c + \frac{t^{\rm lin}}{2t^{\rm quad}}\right) \exp\left(-\frac{1}{2\sigma^2}(x-d)^2\right) \\ & + \int_{t_{\min}}^{\infty} dx\ \delta\left(c - \left(\frac{-t^{\rm lin} + \sqrt{(t^{\rm lin})^2 - 4t^{\rm quad}(t^{\rm SM} - x)}}{2t^{\rm quad}}\right)\right) \exp\left(-\frac{1}{2\sigma^2}(x-d)^2\right) \\ & + \int_{t_{\min}}^{\infty} dx\ \delta\left(c - \left(\frac{-t^{\rm lin} - \sqrt{(t^{\rm lin})^2 - 4t^{\rm quad}(t^{\rm SM} - x)}}{2t^{\rm quad}}\right)\right) \exp\left(-\frac{1}{2\sigma^2}(x-d)^2\right). \end{aligned} \tag{4.32}$$

Simplifying the delta functions in the second and third integrals, we find:

$$P_{c^{(i)}}(c) \propto \delta\left(c + \frac{t^{\rm lin}}{2t^{\rm quad}}\right) \int_{-\infty}^{t_{\min}} dx\ \exp\left(-\frac{1}{2\sigma^2}(x-d)^2\right) + |2ct^{\rm quad} + t^{\rm lin}| \exp\left(-\frac{1}{2\sigma^2}(d - t(c))^2\right)\,. \tag{4.33}$$

This result is different from the posterior distribution obtained by the Bayesian method in Eq. (4.26). Notable features of the posterior distribution $P_{c^{(i)}}(c)$ are: (i) the distribution has a Dirac-delta peak at $c = -t^{\rm lin}/2t^{\rm quad}$; (ii) elsewhere, the distribution is given by the Bayesian posterior distribution rescaled by a prefactor dependent on c. Therefore, the

Monte Carlo replica method will not in general reproduce the Bayesian posterior.

However, one can note that in an appropriate limit, the Bayesian posterior *is* indeed recovered. In particular, suppose that the quadratic EFT cross-section is subdominant compared to the linear term, $t^{\rm lin} \gg t^{\rm quad}$; in this case, we have that $t_{\rm min} \to -\infty$ so that the first term in Eq. (4.33) vanishes and the prefactor of the second term can be approximated with $2ct^{\rm quad} + t^{\rm lin} \approx t^{\rm lin}$. Thus, the Bayesian posterior from Eq. (4.26) is indeed recovered, and we see that the two methods are formally identical for a linear EFT analysis. Further, it is possible to show analytically that for multiple SMEFT parameters and multiple correlated data points, if only linear theory is used the two distributions agree exactly.

This calculation demonstrates that, for quadratic EFT fits, the Monte-Carlo replica method will not in general reproduce the Bayesian posteriors that one would obtain from, say, a nested sampling approach; agreement will only occur provided quadratic EFT corrections are sufficiently subdominant in comparison with the linear ones. For this reason, in this work we restrict the SMEFT-PDF fits based on SIMUNET (which rely on the use of the Monte-Carlo replica method) to linear EFT calculations; we defer the further investigation of the use of the Monte-Carlo replica method, and how it might be modified for use in SIMUNET, to future works. Some preliminary results are presented in Chapter 6.

4.7.2 Application to one-parameter fits

As demonstrated above, the Monte-Carlo replica method will lead to posterior distributions differing from their Bayesian counterparts whenever quadratic EFT corrections dominate over linear ones. Here, we show the numerical impact of these differences in a model case, namely the one-parameter fit of the coefficient c_{dt}^8 to the CMS 13 TeV $t\bar{t}$ invariant mass distribution measurement based on the ℓ+jets final-state [217]. Fig. 4.20 compares the experimental data from this measurement with the corresponding SM theory calculations at NNLO using the NNPDF4.0 (no top) PDF set as input. We observe that the SM theory predictions overshoot the data, especially in the high $m_{t\bar{t}}$ regions, where energy-growing effects enhance the EFT corrections. Given that the pseudodata replicas $d^{(i)}$ are fluctuated around the central value d, the configuration where the SM overshoots the data potentially enhances the contribution of the upper solution in Eq. (4.29) leading to the Dirac delta peak in the posterior Eq. (4.33).

For the case of the c_{dt}^8 coefficient, the quadratic EFT corrections dominate over the linear ones and hence the net effect of a non-zero coefficient is typically an upwards shift of the theory prediction. Indeed, we have verified that for this coefficient the biggest negative correction one can obtain is of order $\sim 2\%$. For the last $m_{t\bar{t}}$ bin, the minimum of the theory cross section $t_{\rm min}$ in Eq. (4.30) is obtained for a value $c_{dt}^8 \approx -0.2$, while for the second to last bin instead $t_{\rm min}$ is minimised by $c_{dt}^8 \approx -0.3$. The combination of

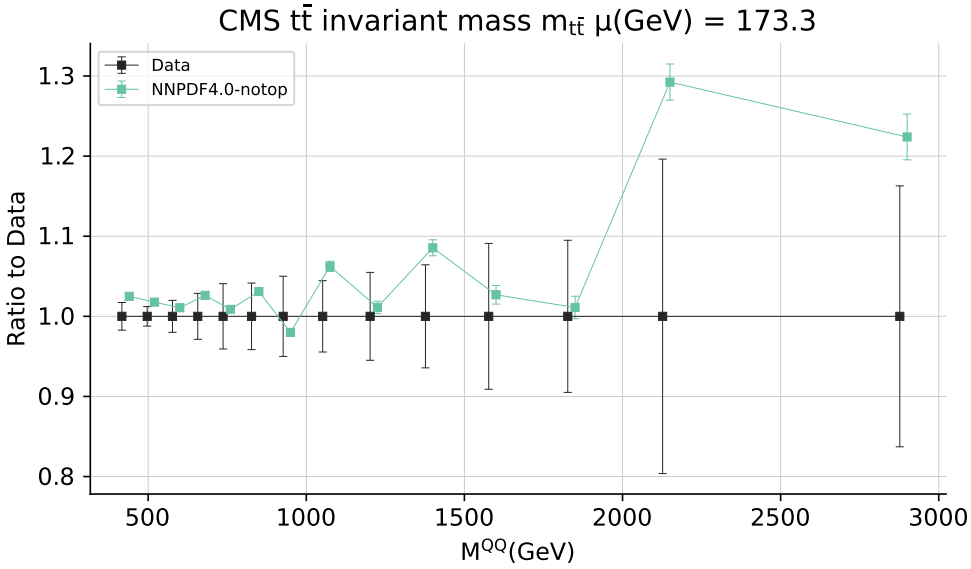


Figure 4.20: Comparison between the experimental data for the top-quark pair invariant mass $m_{t\bar{t}}$ distribution from the ℓ+jets CMS measurement at 13 TeV [217], with the corresponding SM theory calculations at NNLO using the NNPDF4.0 (no top) PDF set as input. For the latter, the error band indicates the PDF uncertainties, and for the former the diagonal entries of the experimental covariance matrix. Results are shown as ratios to the central value of the data. The SM theory predictions overshoot the data, especially in the high $m_{t\bar{t}}$ regions, where energy-growing effects enhance the EFT corrections.

these two features (a dominant quadratic EFT term, and a SM prediction overshooting the data) suggests that the Monte-Carlo replica method's posterior will be enhanced for $c_{dt}^8 \in (-0.3, -0.2)$ as compared to the Bayesian posterior.

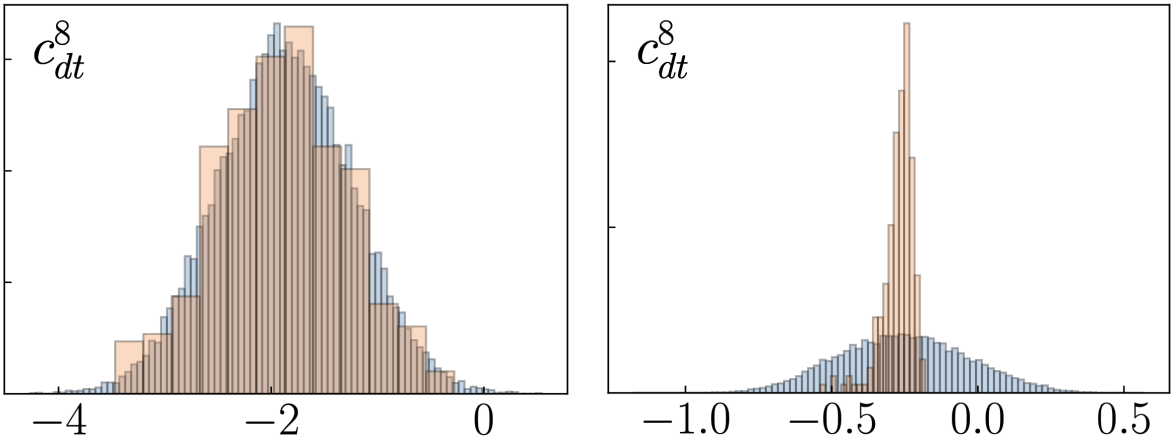


Figure 4.21: Posterior distributions for a one-parameter fit of the four-fermion coefficient c_{dt}^8 with the sole experimental input being the CMS $m_{t\bar{t}}$ distribution displayed in Fig. 4.20. Results are obtained with SMEFiTand we compare the outcome of Nested Sampling (in blue) with that of MCfit (in red) for linear (left panel) and quadratic (right panel) EFT fits.

In Fig. 4.21 we compare the posterior distributions for a one-parameter fit of the four-fermion coefficient c_{dt}^8 with the sole experimental input being the CMS $m_{t\bar{t}}$ distribution displayed in Fig. 4.20. Results are obtained with SMEFiT and we compare the outcome of Nested Sampling (in blue) with that of MCfit (in red) for linear and quadratic EFT fits. The agreement in the linear fit is lost for its quadratic counterpart, with the main difference being a sharp peak in the region $c_{dt}^8 \in (-0.3, -0.2)$ in which the contribution from the delta function in Eq. 4.33 is most marked.

The scenario displayed in Fig. 4.21 is chosen to display the maximum effect, based on a single coefficient with large quadratic EFT corrections, and a dataset where the SM overshoots the data in the $m_{t\bar{t}}$ region where EFT effects are the largest. Within a global fit, these differences are *partially* washed out (indeed the Bayesian and MCfit posterior distributions mostly agree well for the quadratic SMEFiTanalysis, as shown in [43], for the majority of fitted coefficients). Nevertheless, at least in its current implementation, Fig. 4.21 highlights that the Monte-Carlo replica method is affected by pitfalls that prevent its straightforward application to global EFT interpretations of experimental data which include quadratic corrections. We will revisit this more carefully in Chapter 6.

Part II

Future considerations for fitting parton distributions

Chapter 5

Disentangling New Physics effects and parton distributions

[This chapter is based on Ref. [295], worked on in collaboration with Elie Hammou, Zahari Kassabov, Maeve Madigan, Michelangelo Mangano, Luca Mantani, Manuel Morales Alvarado and Maria Ubiali. My contribution to this work comprised: writing the additions to the code which produces the contaminated fits; running various contaminated fits and testing the random seed dependence of the contaminated fits; writing the analysis code which produces BSM bounds based on the contaminated fits; running the $pp \to ZH$ analysis, which is one of those used in Sect. 5.3; running the weighted fits mentioned in Sect. 5.4 as an attempt to disentangle PDFs and New Physics.]

So far, this thesis has exclusively focussed on the simultaneous determination of PDFs and BSM couplings in a selection of models (namely a dark matter model, and two SMEFT scenarios). In this chapter, we ask a more general question: what kind of errors are committed if we do *not* perform a simultaneous extraction, instead fitting the PDFs *assuming* the SM, but New Physics (NP) *is* indeed present in the data that goes into the PDF fit? This could bias the PDFs, and result in further errors if the PDFs are subsequently used in fits of BSM theories themselves.

In order to facilitate this, we work in a setting in which we pretend we know the underlying law of nature, the so-called 'closure test' setting, which we shall describe in significantly more detail in Sect. 5.1. We work in two specific scenarios, based on two possible UV-complete extensions of the SM (which we treat using an EFT approach for convenience), described in Sect. 5.2. In Sect. 5.3, we investigate the impact that these two extensions of the SM have when contaminated data is included in PDF fits, for various strengths of the New Physics. We also assess the effect of using the 'contaminated' PDFs in predictions for other observables, not included in the fit. Finally, in Sect. 5.4, we explore strategies that could be used to disentangle New Physics and PDFs.

5.1 Methodology

In order to systematically study contamination effects from new physics (NP) in PDF fits, we work in a setting in which we pretend we know the underlying law of Nature. In our case the law of Nature consists of the 'true' PDFs, which are low-energy quantities that have nothing to do with new physics, and the 'true' Lagrangian of Nature, which at low energy is well approximated by the Standard Model (SM) Lagrangian but to which we add some heavy new particles. We use these assumptions to generate the artificial data that enter our analysis. We inject the effect of the new particles that we introduce in the Lagrangian in the articifial MC data, and their effect will be visible in some high-energy distributions depending on the underlying model.

The methodology we use throughout this chapter is based on the NNPDF *closure test* framework, first introduced in Ref. [37], and explained in more detail in Ref. [296]. This method was developed in order to assess the quality and the robustness of the NNPDF fitting methodology; in brief it follows three basic steps: (i) assume that Nature's PDFs are given by some fixed reference set; (ii) generate artificial MC data based on this assumption, which we term *pseudodata*; (iii) fit PDFs to the pseudodata using the NNPDF methodology. Various statistical estimators, described in Ref. [296], can then be applied to check the quality of the fit (in broad terms, assessing its difference from the true PDFs), hence verifying the accuracy of the fitting methodology.

In our study, the closure test methodology is adapted to account for the fact that the true theory of Nature may *not* be the SM. In the first step of the closure test methodology, we now assume the existence of some NP model which modifies the MC data generated from the fixed reference PDF set. We investigate the danger posed by subsequently fitting this 'contaminated' pseudodata assuming that the SM is the true theory of Nature.

In this Section, we describe this adapted closure test methodology in more detail. In Sect. 5.1.1, we carefully define the terms *baseline fit* and *contaminated fit*, which shall be used throughout this paper. For completeness of this chapter, we then briefly remind the reader of the salient features of the NNPDF fitting methodology, in particular discussing Monte-Carlo error propagation (of which we of course had much to say in Sect. 4.7, and subsequently will say much more about in Chapter 6). In Sect. 5.1.2, we provide more details on how the MC data is generated in this work. Finally, in Sect 5.1.3 we give an overview of the types of analysis we perform on the fits we obtain.

5.1.1 Basic definitions and fitting methodology

Let us suppose that the true theory of Nature is given by the SM, plus some NP contributions. Under this assumption, observables $T \equiv T(\theta_{\rm SM}, \theta_{\rm NP})$ have a dependence on both the SM parameters $\theta_{\rm SM}$ (in this chapter, exclusively the PDFs), and the NP parameters

$\theta_{\rm NP}$ (for example, masses and couplings of new particles). Let us further fix notation by writing the true values of the parameters $\theta_{\rm SM}, \theta_{\rm NP}$ as $\theta^*_{\rm SM}, \theta^*_{\rm NP}$ respectively; for convenience, we shall also write the true value of the observable T as $T^* \equiv T(\theta^*_{\rm SM}, \theta^*_{\rm NP})$.

Suppose that we wish to perform a fit of some of the theory parameters using experimental measurements of $N_{\rm obs}$ observables, which we package as a single vector $T = (T_1, T_2, ..., T_{N_{\rm obs}})$. All measurements are subject to random observational noise; we assume that this results in the observed data being distributed according to:

$$D_0 = T^* + \eta, \tag{5.1}$$

where η is drawn from the multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, with Σ the experimental covariance matrix describing the correlations between measurements. In the context of the NNPDF closure test methodology [37], the true values of the observables T^* are referred to as *level 0 pseudodata* (L0), whilst the fluctuated values D_0 are referred to as *level 1 pseudodata* (L1).

Once we have generated a sample D_0 of L1 pseudodata, we may perform a fit of some of the theory parameters to this pseudodata. In this work, we shall perform various types of fits with different choices of $\theta^*_{\rm SM}, \theta^*_{\rm NP}$, and different choices of which parameters we are fitting. We define the types of fits as follows:

(1) **Baseline fit.** If there is no new physics, $\theta^*_{\rm NP} \equiv 0$, then the SM is the true theory of Nature. We generate L1 pseudodata D_0 according to the SM. If we subsequently fit the parameters $\theta_{\rm SM}$, we say that we are performing a *baseline fit*. This is precisely equivalent to performing a standard NNPDF closure test.

(2) **Contaminated fit.** If new physics exists, $\theta^*_{\rm NP} \neq 0$, then the SM is *not* the true theory of Nature. We generate L1 pseudodata D_0 according to the SM plus the NP contribution. If we subsequently *only* fit the parameters $\theta_{\rm SM}$ whilst ignoring the NP parameters $\theta_{\rm NP}$, we say that we are performing a *contaminated fit*.

(3) **Simultaneous fit.** If new physics exists, $\theta^*_{\rm NP} \neq 0$, we again generate L1 pseudodata D_0 according to the SM plus the NP contribution. If we subsequently fit *both* the parameters $\theta_{\rm SM}$ *and* $\theta_{\rm NP}$, we say that we are performing a *simultaneous fit*. A closure test of this type is performed in Ref. [103] in order to benchmark the `SIMUnet` methodology. However, we do not perform such fits in this chapter (they were of course performed in Chapter 3 and 4), with our main goal being to assess the possible deficiencies associated with performing *contaminated* fits.

A summary of the first two definitions is given for convenient reference in Table 5.1.

Throughout this chapter, we shall perform only *baseline* and *contaminated* fits; that is, we shall only fit SM parameters, but we shall fit them to pseudodata generated either

Fit name	Nature	Fitted parameters
Baseline	Standard Model: $\theta^*_{\rm NP} \equiv 0$	Standard Model only: $\theta_{\rm SM}$
Contaminated	new physics: $\theta^*_{\rm NP} \neq 0$	Standard Model only: $\theta_{\rm SM}$

Table 5.1: A summary of the definitions of *baseline* and *contaminated* fits used throughout this work.

assuming the law of Nature is given by the SM only, or that it is given by the SM plus some NP contribution.

The NNPDF methodology makes use of the Monte-Carlo (MC) replica method to propagate errors to the PDFs. This involves the generation of additional layer of pseudodata, referred to as *level 2 pseudodata* (L2). Given a L1 pseudodata sample D_0, we generate L2 pseudodata by augmenting D_0 with further random noise ϵ:

$$D = D_0 + \epsilon = T^* + \eta + \epsilon, \tag{5.2}$$

where ϵ is an independent multivariate Gaussian variable, distributed according to $\epsilon \sim \mathcal{N}(0, \Sigma)$, with Σ the experimental covariance matrix. Whilst the L1 pseudodata is sampled only once, the L2 pseudodata D is sampled $N_{\rm rep}$ times, and the best-fit PDFs are obtained to each of the L2 pseudodata samples. This provides an ensemble of PDFs from which statistical estimators, in particular uncertainty bands, can be constructed.

5.1.2 Pseudodata generation

As described above, we assume that the true theory of Nature is the SM plus some new physics. More specifically, in this chapter we take the 'true SM' to mean SM perturbation theory to NNLO QCD accuracy. The true PDF set which shall be used throughout this work is the NNPDF4.0 set [297] (in principle, we are of course allowed to choose any PDF set).

On the other hand, in this chapter we additionally inject two different NP signals. In each case, we base our NP scenario on specific UV-complete models. Furthermore, we choose NP scenarios which are characterised by scales much higher than the energy scales probed by the data, which allows us to justify matching the UV-complete models to a SMEFT parametrisation. The advantage of this approach is that theory predictions become polynomial in the SMEFT Wilson coefficients, which is not necessarily the case in UV-complete models; this allows us maximum flexibility to trial many different values for the 'true' NP parameters. To justify the SMEFT approximation, in Sect. 5.2 we study the validity of the EFT in each case, checking that we only use values of the SMEFT Wilson coefficients which provide good agreement with the UV theory at the linear or quadratic levels, as appropriate.

We also make a K-factor approximation (the validity of which is explicitly checked in Ref. [298] in the case of the $\hat{W}, \hat{Y}$ parameter scenarios) to avoid expensive computation of fast interpolation grids for the PDFs. As a result, the formula for the 'true' value of an observable takes the schematic form:

$$T \equiv \left(1 + cK_{\text{lin}} + c^2 K_{\text{quad}}\right) \hat{\sigma}^{\text{SM}} \otimes \mathcal{L}, \tag{5.3}$$

where $\mathcal{L}$ denotes either the PDFs or PDF luminosities for NNPDF4.0 (depending on whether the observable is a deep inelastic scattering or hadronic observable), c denotes the SMEFT Wilson coefficient(s) under consideration, $\hat{\sigma}^{\text{SM}}$ is the SM partonic cross-section computed at NNLO in QCD perturbation theory, and $K_{\text{lin}}, K_{\text{quad}}$ are the SMEFT K-factors.

5.1.3 Post-fit analysis

Once we have produced a contaminated fit, where PDFs have been fitted using SM theory to data produced with the SM plus some NP contribution, several natural questions arise.

Detection of contamination. Is it possible to detect contamination of the PDF fit by the NP effects? If there are many datasets entering the fit which are *not* affected by NP, it might be the case that datasets which *are* affected by NP could appear inconsistent, and might be poorly described by the resulting fit.

In order to address this point, we use the NNPDF dataset selection criteria, discussed in detail in Ref. [297]. We consider both the χ^2-statistic of the resulting contaminated PDF fit to each dataset entering the fit, and also consider the number of standard deviations

$$n_\sigma = (\chi^2 - n_{\text{dat}})/\sqrt{2n_{\text{dat}}} \tag{5.4}$$

of the χ^2-statistic from the expected χ^2 for each dataset. If $\chi^2/n_{\text{dat}} > 1.5$ and $n_\sigma > 2$ for a particular dataset, the dataset would be flagged by the NNPDF selection criteria, indicating an inconsistency with the other data entering the fit.

There are two possible outcomes of performing such a dataset selection analysis on a contaminated fit. In the first instance, the datasets affected by NP are flagged by the dataset selection criterion. If a dataset is flagged according to this condition, then a weighted fit is performed, *i.e.* a fit in which a dataset is given a larger weight inversely proportional to the number of datapoints, effected by modifying the χ^2-statistic to:

$$\chi^2_w = \frac{1}{n_{\text{dat}} - n^{(j)}_{\text{dat}}} \sum_{i=1, i\neq j}^{n_{\text{exp}}} n^{(i)}_{\text{dat}} \chi^2_i \; + \; w^{(j)} \chi^2_j, \tag{5.5}$$

where χ^2_i is the χ^2-statistic evaluated on the ith dataset, and:

$$w^{(j)} = n_{\rm dat}/n^{(j)}_{\rm dat}. \tag{5.6}$$

If the data-theory agreement improves by setting it below thresholds and the data-theory agreement of the other datasets does not deteriorate in any statistically relevant way, then the dataset is kept, else the dataset is discarded, on the basis of the inconsistency with the remaining datasets. In the second instance, the datasets are *not* flagged, and are consistent enough that the contaminated fit would pass undetected as a *bona fide* SM PDF fit. We introduce the following terms to describe each of these cases: in the former case, we say that the PDF was *unable to absorb* the NP; in the latter case, we say that the PDF has *absorbed* the NP.

Distortion of NP bounds. Can using a contaminated fit in a subsequent fit of NP effects lead to misleading bounds? In more detail, suppose that we construct a contaminated fit which has absorbed NP - that is, the contamination would go undetected by the NNPDF dataset selection criterion. In this case, we would trust that our contaminated fit was a perfectly consistent SM PDF fit, and might try to use it to subsequently fit the underlying parameters in the NP scenario.

There are two possible outcomes of such a fit. The contamination of the PDFs may be weak enough for the NP bounds that we obtain to be perfectly sensible, containing the true values of the NP parameters. On the other hand, the absorption of the NP may be strong enough for the NP bounds to be distorted, no longer capturing the true underlying values of the NP parameters. The second case is particularly concerning, and if it can occur, points to a clear need to disentangle PDFs and possible NP effects.

Distortion of SM predictions. Finally, we ask: can using a contaminated fit lead to poor agreement on new datasets which are not affected by NP? In particular, suppose that we are again in the case where NP has been absorbed by a contaminated fit, so that the NP signal has gone undetected. If we were to use this contaminated fit to make predictions for an observable which is *not* affected by the NP, it is interesting to see whether the data is well-described or not; if the contamination is sufficiently strong, it may appear that the dataset is inconsistent with the SM. This could provide a route for disentangling PDFs and NP.

5.2 New Physics scenarios

As discussed in Sect. 5.1, throughout this chapter we have assumed that the theory of Nature is the SM plus some new physics, and generated pseudodata accordingly. In

this section, following the methodology presented in Sect. 5.1.2, we extend the SM by UV-complete models by introducing heavy new fields.

We choose simple extensions of the SM which correspond to 'universal theories' [299], the effect of which on our dataset can be well-described in an EFT approximation using the oblique corrections $\hat{Y}$ and $\hat{W}$ [121, 122, 123], which should be familiar from Sect. 3.3. We consider the following scenarios:

- **Scenario I:** A flavour universal Z', which could be associated to an additional $U(1)_Y$ gauge symmetry. We give a mass to the field assuming it is generated by some higher energy physics. It corresponds to a new heavy neutral bosonic particle. At the EFT level, the effect of this model on our dataset can be described by the $\hat{Y}$ parameter.

- **Scenario II:** A flavour universal W' charged under $SU(2)_L$. Once again, we directly add a mass term to the Lagrangian. This corresponds to two new heavy charged bosonic particles (W'^+ and W'^-) as well as a new heavy neutral boson, similar to the Z' but which only couples to left-handed particles. At the EFT level, the effect of this model on our dataset can be described by the $\hat{W}$ parameter.

The following subsections are devoted to describing each of these NP scenarios. In particular, in each case we use tree-level matching to obtain a parametrisation of the model in terms of dimension 6 operators of the SMEFT, making use of the matching provided in Ref. [300] to do so. We identify the observables in our dataset affected by each NP scenario, and in each case we compare the UV and EFT predictions. Finally, we identify values of the model parameters for which the EFT description is justified at the projected luminosity of the HL-LHC, and for which existing constraints on the models are avoided.

5.2.1 Scenario I: A flavour-universal Z' model

The addition to the SM of a new spin-1 boson Z' associated to a new gauge symmetry $U(1)_Y$, a mass $M_{Z'}$ and a coupling coefficient $g_{Z'}$ yields the following Lagrangian:

$$\begin{aligned}\mathcal{L}_{\text{UV}}^{Z'} = \mathcal{L}_{\text{SM}} &- \frac{1}{4} Z'_{\mu\nu} Z'^{\mu\nu} + \frac{1}{2} M_{Z'}^2 Z'_\mu Z'^\mu \\ &- g_{Z'} Z'_\mu \sum_f Y_f \bar{f} \gamma^\mu f - Y_\varphi g_{Z'} (Z'_\mu \varphi^\dagger i D^\mu \varphi + \text{h.c.}) \,.\end{aligned} \tag{5.7}$$

We sum the interactions with the fermions $f \in \{q, u, d, \ell, e\}$, where Y_f is the corresponding hypercharge: $(Y_q, Y_u, Y_d, Y_l, Y_e, Y_\varphi) = (\frac{1}{6}, \frac{2}{3}, -\frac{1}{3}, -\frac{1}{2}, -1, \frac{1}{2})$. The kinetic term is given by $Z'_{\mu\nu} = \partial_\mu Z'_\nu - \partial_\nu Z'_\mu$. The covariant derivative is $D_\mu = \partial_\mu + \frac{1}{2} i g \sigma^a W_\mu^a + i g' Y_\varphi B_\mu + i g_{Z'} Y_\varphi Z'_\mu$. We neglect the mixing between the Z' and the SM gauge bosons. Note that quark and lepton flavour indices are suppressed, and that the couplings to quarks and leptons are

flavour diagonal. The new gauge interaction is anomaly-free, as it has the same hypercharge-dependent couplings to fermions as the SM fields [301]. Models of Z' bosons and their impact on LHC data have been widely studied; see for example Refs. [302, 303, 304, 305].

Bosonic	$\mathcal{O}_{\varphi D}$, $\mathcal{O}_{\varphi\Box}$, $\mathcal{O}_{\varphi l}^{(1)}$, $\mathcal{O}_{\varphi q}^{(1)}$, $\mathcal{O}_{\varphi e}$, $\mathcal{O}_{\varphi u}$, $\mathcal{O}_{\varphi d}$
4-fermion $(\bar{L}L)(\bar{L}L)$	$\mathcal{O}_{ll}$, $\mathcal{O}_{qq}^{(1)}$, $\mathcal{O}_{lq}^{(1)}$
4-fermion $(\bar{R}R)(\bar{R}R)$	$\mathcal{O}_{ee}$, $\mathcal{O}_{uu}$, $\mathcal{O}_{dd}$, $\mathcal{O}_{ed}$, $\mathcal{O}_{eu}$, $\mathcal{O}_{ud}^{(1)}$
4-fermion $(\bar{L}L)(\bar{R}R)$	$\mathcal{O}_{le}$, $\mathcal{O}_{ld}$, $\mathcal{O}_{lu}$, $\mathcal{O}_{qu}^{(1)}$, $\mathcal{O}_{qd}^{(1)}$

Table 5.2: Warsaw basis operators generated by the Z' model of Eq. (5.7).

Tree-level matching of $\mathcal{L}_{\rm UV}^{Z'}$ to the dimension 6 SMEFT produces the Warsaw basis [300] operators in Table 5.2. The exhaustive SMEFT Lagrangian is given by:

$$\begin{aligned}\mathcal{L}_{\rm SMEFT}^{Z'} = \mathcal{L}_{\rm SM} - \frac{g_{Z'}^2}{M_{Z'}^2}\Big(&2Y_\varphi^2\mathcal{O}_{\varphi D} + \frac{1}{2}Y_\varphi^2\mathcal{O}_{\varphi\Box} \\ &+ Y_\varphi Y_l\mathcal{O}_{\varphi l}^{(1)} + Y_\varphi Y_q\mathcal{O}_{\varphi q}^{(1)} + Y_\varphi Y_e\mathcal{O}_{\varphi e} + Y_\varphi Y_d\mathcal{O}_{\varphi d} + Y_\varphi Y_u\mathcal{O}_{\varphi u} \\ &+ \frac{1}{2}Y_l^2\mathcal{O}_{ll} + \frac{1}{2}Y_q^2\mathcal{O}_{qq}^{(1)} + Y_qY_l\mathcal{O}_{lq}^{(1)} \\ &+ \frac{1}{2}Y_e^2\mathcal{O}_{ee} + \frac{1}{2}Y_u^2\mathcal{O}_{uu} + \frac{1}{2}Y_d^2\mathcal{O}_{dd} + Y_eY_d\mathcal{O}_{ed} + Y_eY_u\mathcal{O}_{eu} + Y_uY_d\mathcal{O}_{ud}^{(1)} \\ &+ Y_eY_l\mathcal{O}_{le} + Y_uY_l\mathcal{O}_{lu} + Y_dY_l\mathcal{O}_{ld} + Y_eY_q\mathcal{O}_{qe} + Y_uY_q\mathcal{O}_{qu}^{(1)} + Y_dY_q\mathcal{O}_{qd}^{(1)}\Big)\,.\end{aligned} \tag{5.8}$$

The leading effect of this model on the data entering our analysis is to modify the Drell-Yan and Deep Inelastic Scattering datasets; in particular the high-mass neutral current Drell-Yan tails [298] will be affected. The Z' may have an additional impact on top quark and dijet data through four-quark operators such as $\mathcal{O}_{qq}^{(1)}$, however the effect is negligible and we do not consider it here.

The effect of the Z' on high-mass Drell-Yan is dominated by the energy-growing four-fermion operators [120, 306, 305]. By neglecting the subdominant operators involving the Higgs doublet φ, we can add the operators of the last three lines of Eq. (5.8) in the following way:

$$\mathcal{L}_{\rm SMEFT}^{Z'} = \mathcal{L}_{\rm SM} - \frac{g_{Z'}^2}{2M_{Z'}^2}J_Y^\mu J_{Y,\mu}, \qquad J_Y^\mu = \sum_f Y_f\bar{f}\gamma^\mu f\,. \tag{5.9}$$

We can describe the new physics introduced in this type of scenario with the $\hat{Y}$ parameter (which should be familiar from Sect. 3.3):

$$\mathcal{L}_{\rm SMEFT}^{Z'} = \mathcal{L}_{\rm SM} - \frac{g'^2\hat{Y}}{2m_W^2}J_Y^\mu J_{Y,\mu}, \qquad \hat{Y} = \frac{g_{Z'}^2}{M_{Z'}^2}\frac{m_W^2}{g'^2}\,. \tag{5.10}$$

The $\hat{Y}$ parameters allows us to write the Lagrangian using SM parameters. We can write the relation between $\hat{Y}$ and the Z' parameters $g_{Z'}$, $M_{Z'}$ as follows:

$$\frac{g_{Z'}^2}{M_{Z'}^2} = 4\sqrt{2}G_F\hat{Y}\left(\frac{m_Z^2 - m_W^2}{m_W^2}\right), \tag{5.11}$$

where we make use of the $\{m_W, m_Z, G_F\}$ electroweak input scheme, and take the following as input parameters:

$$G_F = 1.16639 \times 10^{-5}\ \text{GeV}, \qquad m_W = 80.352\ \text{GeV}, \qquad m_Z = 91.1876\ \text{GeV}. \tag{5.12}$$

In Fig. 5.1 we compare the predictions of the UV-complete Z' model and the corresponding EFT parametrisation for differential cross-sections of the Drell-Yan processes $pp \to \ell^+\ell^-$. The predictions are computed assuming $\sqrt{s} = 14$ TeV, using `MadGraph5_aMC@NLO`. We compare the SM, the full UV-complete model, the linear-only $\mathcal{O}(\Lambda^{-2})$ EFT and linear-plus-quadratic $\mathcal{O}(\Lambda^{-4})$ EFT predictions assuming $g_{Z'} = 1$, for three benchmark values of the Z' mass: $M_{Z'} = 14.5$ TeV, $M_{Z'} = 18.7$ TeV and $M_{Z'} = 32.5$ TeV. Such large values of $M_{Z'}$ are clearly well beyond the possible direct reach of direct Z' searches at ATLAS and CMS [307].

In the top panel we plot the differential cross-section with respect to the dilepton invariant mass, in the middle panel we plot the ratio of the full Z' model to the SM, and in the lower panel we plot the ratio of the EFT to the full Z' model predictions. First we observe that the UV model predictions differ from the SM predictions in a way that could be measurable at the HL-LHC. In the lower panels, we observe the point at which the linear EFT corrections fail to describe the UV physics, and the quadratic EFT contributions begin to become non-negligible; in the same way, the quadratic dimension 6 EFT description starts failing when the dimension 8 SMEFT operators become important [308].

As displayed in the top right panel of Fig. 5.1, even if for $M_{Z'} = 14.5$ TeV the linear EFT corrections start failing describing the UV models at high energies, from $M_{Z'} = 18.7$ TeV onward, the linear EFT describes the UV physics faithfully for dilepton invariant masses up to 4 TeV. The deviations from new physics are over 10% for $M_{Z'} = 18.7$ and $M_{Z'} = 14.5$ TeV. We will implement our PDF 'contamination' working in this area of the parameter space, and making use of linear EFT corrections only.

Finally, we note that we have compared the SMEFT predictions with those obtained including the SMEFT operators with the Higgs doublet, such as $\mathcal{O}_{\varphi l}^{(1)}$ and $\mathcal{O}_{\varphi e}$. We find that these operators have no visible impact. They are only competitive with the four-fermion corrections at lower energies (around 500 GeV), and at this scale the new physics has very little impact on the SM predictions. When the influence of the heavy new physics starts to be noticeable at higher invariant mass, the four-fermion operators, whose impact grows faster with energy, completely dominate the SMEFT corrections. Thus, we have verified

that our parametrisation in terms of the $\hat{Y}$ parameter reflects the UV physics to a very good degree.

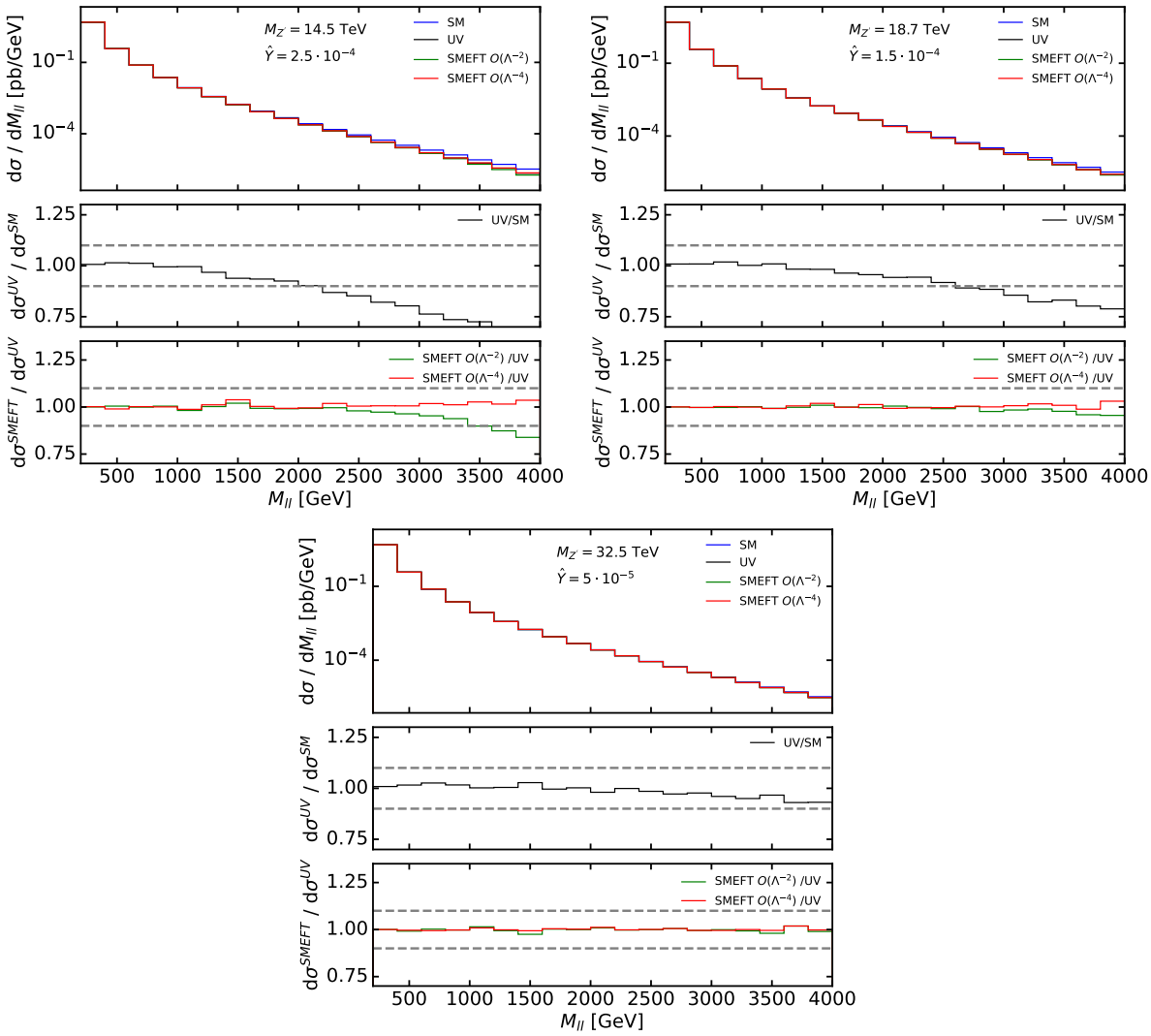


Figure 5.1: Predictions for neutral current Drell-Yan differential cross sections in dilepton invariant mass. We show the SM predictions compared to the predictions for a Z' of different masses corresponding to different $\hat{Y}$ values, assuming $g_{Z'} = 1$. The dashed lines indicate a deviation of 10%, a rough indication of a visible effect. Top left: mass of 14.5 TeV, corresponding to $\hat{Y} = 25 \cdot 10^{-5}$. Top right: mass of 18.7 TeV, corresponding to $\hat{Y} = 15 \cdot 10^{-5}$. Bottom: mass of 32.5 TeV, corresponding to $\hat{Y} = 5 \cdot 10^{-5}$.

5.2.2 Scenario II: A flavour universal W' model

We now consider a new $SU(2)_L$ triplet field $W'^{a,\mu}$, where $a \in \{1, 2, 3\}$ denotes an $SU(2)_L$ index. We add a mass $M_{W'}$, and denote the W' coupling coefficient by $g_{W'}$. Similarly to what happens with the SM W field, the W'^1 and W'^2 components mix to form the W'^+ and W'^- particles, while the W'^3 component gives a neutral boson similar to the

Z' but which only couples to left-handed fields. The model is described by the following Lagrangian:

$$\begin{aligned}\mathcal{L}_{\rm UV}^{W'} = \mathcal{L}_{\rm SM} &- \frac{1}{4} W'^{a}_{\mu\nu} W'^{a,\mu\nu} + \frac{1}{2} M_{W'}^2 W'^{a}_{\mu} W'^{a,\mu} \\ &- g_{W'} W'^{a,\mu} \sum_{f_L} \bar{f}_L T^a \gamma^\mu f_L - g_{W'} (W'^{a,\mu} \varphi^\dagger T^a i D_\mu \varphi + \text{h.c.}) \,,\end{aligned} \tag{5.13}$$

where we sum over the left-handed fermions: $f_L \in \{q, \ell\}$. The $SU(2)_L$ generators are given by $T^a = \frac{1}{2}\sigma^a$ where σ^a are the Pauli matrices. The kinetic term is given by $W'^{a}_{\mu\nu} = \partial_\mu W'^{a}_{\nu} - \partial_\nu W'^{a}_{\mu} - i g_{W'} [W'^{a}_{\mu}, W'^{a}_{\nu}]$. The covariant derivative is given by $D_\mu = \partial_\mu + \frac{1}{2} i g \sigma^a W^a_\mu + i g' Y_\varphi B_\mu$. As above, we neglect the mixing with the SM gauge fields.

Tree-level matching of $\mathcal{L}_{\rm UV}^{W'}$ to the dimension 6 SMEFT produces the Warsaw basis operators in Table 5.3, where we have distinguished the operators $(O_{ll})_{ij} = (l_i \gamma^\mu l_i)(l_j \gamma^\mu l_j)$ and $(O'_{ll})_{ij} = (l_i \gamma^\mu l_j)(l_j \gamma^\mu l_i)$ [113].

Bosonic	$\mathcal{O}_{\varphi\Box}$, $\mathcal{O}_{\varphi}$, $\mathcal{O}^{(3)}_{\varphi l}$, $\mathcal{O}^{(3)}_{\varphi q}$
Yukawa	$\mathcal{O}_{e\varphi}$, $\mathcal{O}_{d\varphi}$, $\mathcal{O}_{u\varphi}$
4-fermion $(\bar{L}L)(\bar{L}L)$	$\mathcal{O}_{ll}$,$\mathcal{O}'_{ll}$, $\mathcal{O}^{(3)}_{qq}$, $\mathcal{O}^{(3)}_{lq}$

Table 5.3: Warsaw basis operators generated by the W' model of Eq. (5.13).

The complete SMEFT Lagrangian is given by [300]:

$$\begin{aligned}\mathcal{L}_{\rm SMEFT}^{W'} = \mathcal{L}_{\rm SM} - \frac{g_W^2}{M_W^2}\Big(&- \frac{1}{8}\mathcal{O}_{ll} + \frac{1}{4}\mathcal{O}'_{ll} + \frac{1}{8}\mathcal{O}^{(3)}_{qq} + \frac{1}{4}\mathcal{O}^{(3)}_{lq} \\ &+ \lambda_\varphi \mathcal{O}_\varphi + \frac{3}{8}\mathcal{O}_{\varphi\Box} + \frac{1}{4}\mathcal{O}^{(3)}_{\varphi q} + \frac{1}{4}\mathcal{O}^{(3)}_{\varphi l} \\ &+ \frac{1}{4}(y_e)_{ij}(\mathcal{O}_{e\varphi})_{ij} + \frac{1}{4}(y_u)_{ij}(\mathcal{O}_{u\varphi})_{ij} + \frac{1}{4}(y_d)_{ij}(\mathcal{O}_{d\varphi})_{ij}\Big) \,.\end{aligned} \tag{5.14}$$

As in the case of the Z', the leading effect of this model on our dataset is to modify the Drell-Yan and Deep Inelastic Scattering datasets; however this time both charged current and neutral current Drell-Yan will be affected. This impact is dominated by the four-fermion interactions in the first line of Eq. (5.14), which sum to:

$$\mathcal{L}_{\rm SMEFT}^{W'} = \mathcal{L}_{\rm SM} - \frac{g_{W'}^2}{2M_{W'}^2} J_L^{a,\mu} J^a_{L,\mu}, \qquad J_L^{a,\mu} = \sum_{f_L} \bar{f}_L T^a \gamma^\mu f_L \,. \tag{5.15}$$

We can describe the new physics introduced in this type of scenario with the $\hat{W}$ parameter

(again, compare with Sect. 3.3):

$$\mathcal{L}^{W'}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} - \frac{g^2 \hat{W}}{2m_W^2} J_L^{a,\mu} J^a_{L,\mu}, \qquad \hat{W} = \frac{g^2_{W'}}{g^2} \frac{m_W^2}{M^2_{W'}} \,. \tag{5.16}$$

Using Fermi's constant, we can write the relation between the UV parameters and $\hat{W}$ in the following way:

$$\frac{g^2_{W'}}{M^2_{W'}} = 4\sqrt{2} G_F \hat{W} \,. \tag{5.17}$$

Again, by fixing $g_{W'} = 1$, each $M_{W'}$ can be associated to a value of $\hat{W}$.

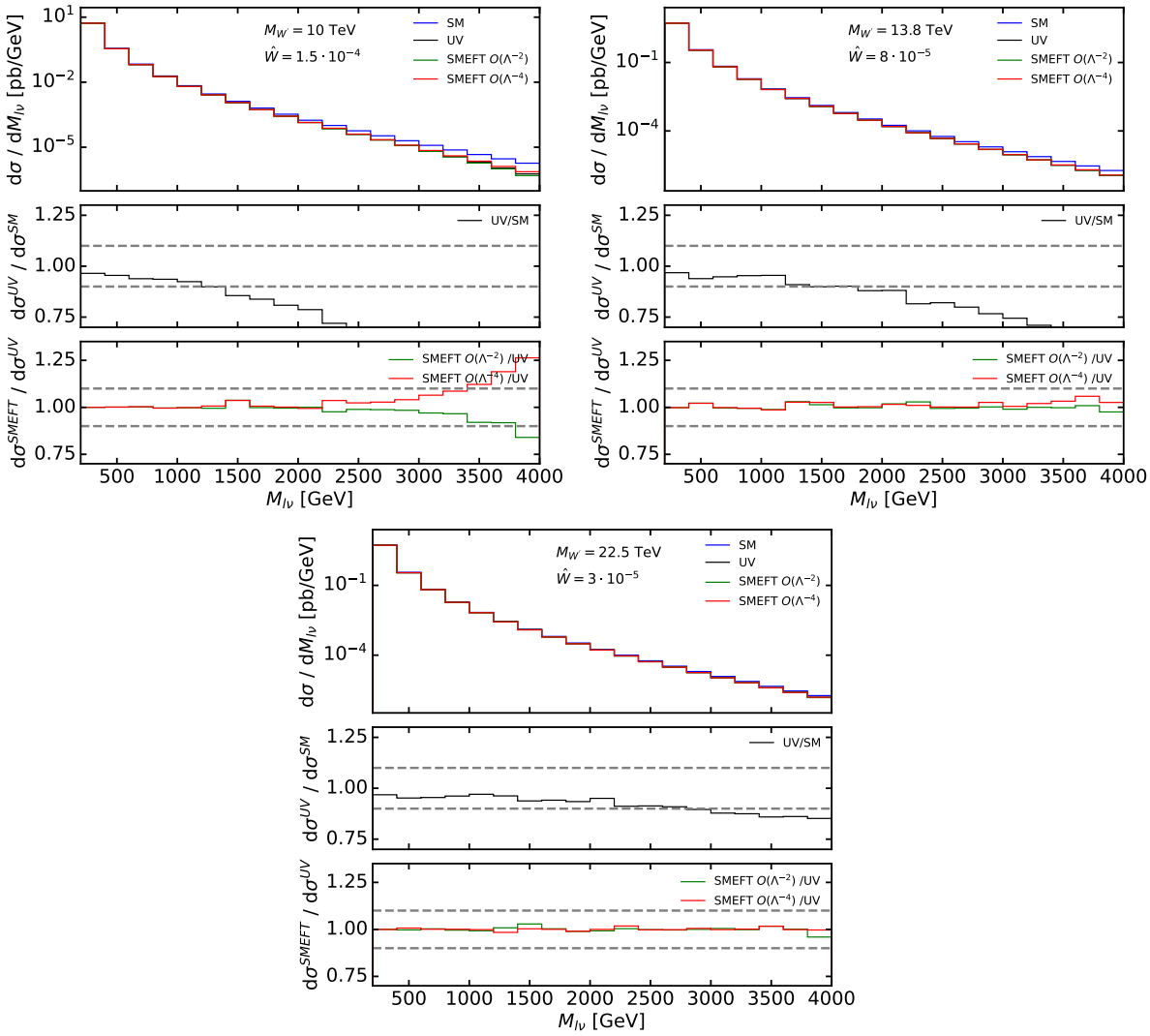


Figure 5.2: Predictions for charged current Drell-Yan ($p\bar{p} \to l^- \bar{\nu}$) differential cross sections in dilepton invariant mass. We show the SM predictions compared to the predictions for a W' of different masses corresponding to different $\hat{W}$ values, assuming $g_{W'} = 1$. Top left: mass of 10 TeV, corresponding to $\hat{W} = 15 \cdot 10^{-5}$. Top right: mass of 13.8 TeV, corresponding to $\hat{W} = 8 \cdot 10^{-5}$. Bottom: mass of 22.5 TeV, corresponding to $\hat{W} = 3 \cdot 10^{-5}$.

In Fig. 5.2 we perform a comparison of the UV-complete W' model and the EFT

predictions. We assess the differences between the EFT parametrisation and the UV model description by studying the charged current Drell-Yan process, $p\bar{p} \to l^- \bar{\nu}$, assuming $g_{W'} = 1$, at three benchmark values of the W' mass: $M_{W'} = 10$ TeV, $M_{W'} = 13.8$ TeV and $M_{W'} = 22.5$ TeV. A similar comparison could be made in neutral current Drell-Yan; however we expect the dominant effect of the W' to occur in charged current Drell-Yan, and therefore this process will provide the leading sensitivity to differences between the UV model and EFT parametrisation. As displayed in the top right panel of Fig. 5.2, even if for $M_{W'} = 10$ TeV the linear EFT corrections start failing describing the UV models at high energies, from $M_{W'} = 13.8$ TeV onward, the linear EFT describes the UV physics faithfully for dilepton invariant masses up to 4 TeV. The deviations from new physics are over 10% for all three cases here. This is the region we will explore for the PDF 'contamination' in the next section.

Finally, our analysis also reveals that the SMEFT operators involving a Higgs doublet φ have no impact on the predictions, for the same reason we presented in the Z' case. This means that this model built with the $\hat{W}$ parameters describes the UV physics faithfully.

5.3 Contamination from Drell-Yan large invariant-mass distributions

In this Section, after presenting the analysis settings in terms of theory predictions and data, we explore in detail the effects of new heavy vector bosons in the high-mass Drell-Yan distribution tails and how fitting this data assuming the SM would modify the data-theory agreement and the PDFs. We will see that in some scenarios the PDFs manage to mimic the effects of new physics in the high tails without deteriorating the data-theory agreement in any visible way. In this cases PDFs can actually 'fit away' the effects of new physics. In the following sections we will explore the phenomenological consequences of using such 'contaminated' PDF sets and we will see that they might significantly distort the interpretation of HL-LHC measurements. Finally we conclude by devising strategies to spot the contamination by including in a PDF fit complementary observables that highlight the incompatibility of the high-mass Drell-Yan tails with the bulk of the data.

5.3.1 Analysis settings

For this analysis we generate a set of artificial Monte Carlo data, which comprises 4771 data points, spanning a broad range of processes. The Monte Carlo data that we generate are either taken from current Run I and Run II LHC data or from HL-LHC projections. The uncertainties in the former category are more realistic, as they are taken from the experimental papers (we remind the reader that the central measurement is generated by

the underlying law of Nature according to Eq. (5.1)), while the uncertainties on projected HL-LHC data are generated according to specific projections.

As far as the current data is concerned, we generate MC data that cover all the observables included in the NNPDF4.0 analysis [297]. In particular, in the category of Drell-Yan, we include the NC Drell-Yan that follow the kinematic distributions and the errors analysed by ATLAS at 7, and 8 TeV [309, 310] and CMS at 7, 8, and 13 TeV [311, 129, 104]. These LHC measurements are not only used to constrain the PDF, but are also sufficiently sensitive to the BSM scenarios considered in Sect. 5.2.

As far as HL-LHC pseudo-data are concerned, we include the high-mass Drell-Yan projections that we produced in Chapters 2 and 3, inspired by the HL-LHC projections studied in Ref. [291]. The invariant mass distribution projections are generated at $\sqrt{s} = 14$ TeV, assuming an integrated luminosity of $\mathcal{L} = 6$ ab^{-1} (3 ab^{-1} collected by ATLAS and 3 ab^{-1} by CMS). Both in the case of NC and CC Drell-Yan cross sections, the MC data were generated using the `MadGraph5_aMCatNLO` NLO Monte Carlo event generator [162] with additional K-factors to include the NNLO QCD and NLO EW corrections. The MC data consist of four datasets (associated with NC/CC distributions with muons/electrons in the final state), each comprising 16 bins in the m_{ll} invariant mass distribution or transverse mass m_T distributions with both m_{ll} and m_T greater than 500 GeV, with the highest energy bins reaching $m_{ll} = 4$ TeV ($m_T = 3.5$ TeV) for NC (CC) data. The rationale behind the choice of number of bins and the width of each bin was outlined in Chapter 2, and stemmed from the requirement that the expected number of events per bin was big enough to ensure the applicability of Gaussian statistics. The choice of binning for the m_{ll} (m_T) distribution at the HL-LHC is displayed in Fig. 3.8.

The kinematic coverage of the data points used in this study are shown in Fig. 5.3. The points are shown in (x, Q^2) space with the data points that are modified by the EFT operators highlighted with a border, such points thus also constrain the Wilson coefficients as well as the PDFs. We note that, although DIS theory predictions are modified by the operators we consider in the two benchmark scenarios, the change in the HERA DIS cross sections upon the variation of the Wilson coefficients under consideration is minimal, as is explicitly assessed in Chapter 3.

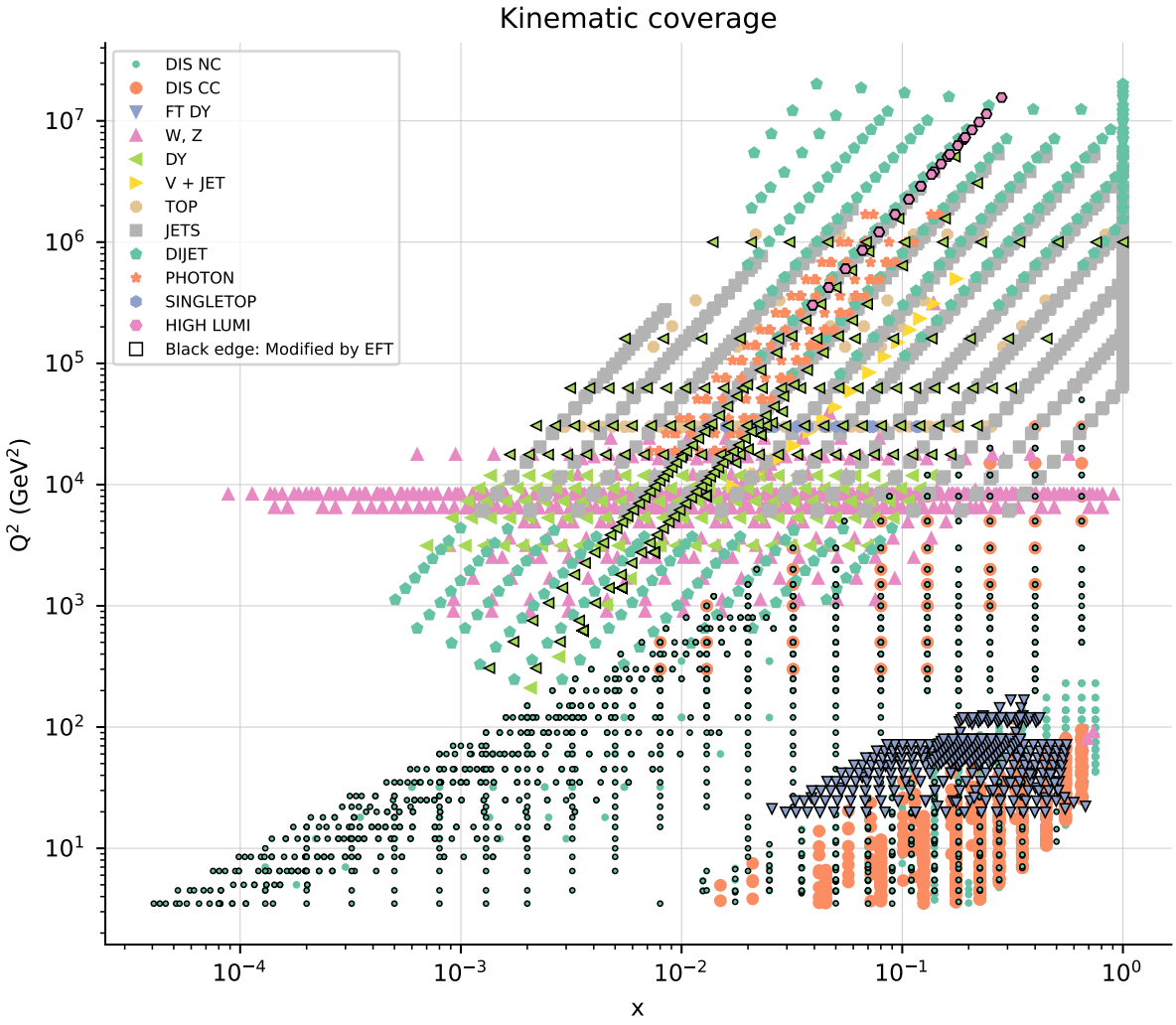


Figure 5.3: Kinematic coverage of data points included in the fit. The EFT corrections for this study have been computed for the points which are highlighted with a black edge. The values of x have been computed using a linear order approximation.

In what follows we will assess the impact of the injection of NP in the data on the fitted PDFs, by looking at the integrated luminosities for the parton pair i, j, which is defined as:

$$\mathcal{L}_{ij}(m_X, \sqrt{s}) = \frac{1}{s} \int_{-y}^{y} d\tilde{y} \left[f_i\left(\frac{m_X}{\sqrt{s}} e^{\tilde{y}}, m_X\right) f_j\left(\frac{m_X}{\sqrt{s}} e^{-\tilde{y}}, m_X\right) + (i \leftrightarrow j) \right], \tag{5.18}$$

where $f_i \equiv f_i(x, Q)$ is the PDF corresponding to the parton flavour i , and the integration limits are defined by:

$$y = \ln\left(\frac{\sqrt{s}}{m_X}\right). \tag{5.19}$$

In particular we will focus on the luminosities that are most constrained by the Neutral

Current (NC) and Charged Current (CC) Drell-Yan data respectively, namely

$$\mathcal{L}^{\rm NC}(m_X, \sqrt{s}) = \mathcal{L}_{u\bar{u}}(m_X, \sqrt{s}) + \mathcal{L}_{d\bar{d}}(m_X, \sqrt{s}), \tag{5.20}$$

$$\mathcal{L}^{\rm CC}(m_X, \sqrt{s}) = \mathcal{L}_{u\bar{d}}(m_X, \sqrt{s}) + \mathcal{L}_{d\bar{u}}(m_X, \sqrt{s}). \tag{5.21}$$

5.3.2 Effects of new heavy bosons in PDF fits

In Fig. 5.4 we display the benchmark points that we consider, corresponding to the two scenarios described in Sect. 5.2. Namely, the points along the vertical axis correspond to the flavour-universal Z' model (Scenario I), while those along the horizontal axis correspond to the flavour-universal W' model (Scenario II). The benchmark points are compared to projected constraints from the HL-LHC. In particular, we consider the most up-to-date constraints from the analysis of a fully-differential Drell-Yan projection in the HL-LHC regime, as given by Ref. [305].

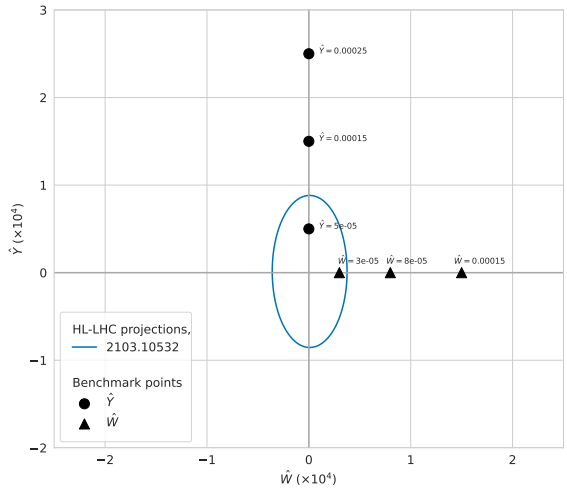


Figure 5.4: Benchmark $\hat{Y}$ and $\hat{W}$ points explored in this analysis compared to the constraints at 95% CL as given by the analysis of fully-differential Drell-Yan projections given in Ref. [305].

In order to estimate the effect of a heavy Z' (W') in Nature and the ability of PDFs to fit it away, we inject new physics in the artificial Monte Carlo data by setting $\hat{Y} \neq 0$ ($\hat{W} \neq 0$) to the values that we consider in our benchmark (see Fig. 5.4) and we measure the effect on the fit quality and on the PDFs. To assess the fit quality, we generate *L1 pseudo-data* as in Eq. (5.1) according to 1000 variations of the random seed and compare the distributions of $\chi^{2(k)}/n_{\rm dat}$ and $n^{(k)}_{\rm sigma}$ across the 1000 random seed (k) variations for the baseline and the 3 benchmark values in each of the two scenarios. If the distributions shift above the critical levels defined in Sect. 5.1, then the PDFs have *not* been able to absorb the effects of new physics and the datasets that display a bad data-theory agreement would

be excluded from a PDF fit. If instead the distributions remain statistically equivalent to those of the baseline PDF fit, then the PDFs have been able to absorb new physics.

Note that in this exercise the distribution across random seed values is calculated by keeping the PDF fixed to the value obtained with a given random seed, while if we were refitting them for each random seed, we would obtain slightly different PDFs. A comparison at the level of PDFs and parton luminosities is then performed to assess whether the absorption of new physics shifts them significantly with respect to the baseline PDFs. We have verified that the effect is negligible and does not modify the results. A more detailed account of the contaminated PDF's random seed dependence is given in App. C. The goal of this exercise is to estimate the maximum strength of new physics effects beyond which PDFs are no longer able to absorb the effect, and subsequently assess whether the effect is significant or not.

(i) Scenario I

In the flavour-universal Z' model we inject three non-zero values of $\hat{Y} = 5 \cdot 10^{-5}$, $15 \cdot 10^{-5}$, $25 \cdot 10^{-5}$. In Fig. 5.5 we display the $\chi^{2(k)}$ and $n_\sigma^{(k)}$ distributions across the 1000 k random seeds for a selection of the datasets included in each of the fits. In particular, we display the datasets in which a shift occurs either because of the direct effect of the non-zero Wilson coefficients in the partonic cross sections (such as the high-mass Drell-Yan in the HL-LHC projections) or because of the indirect effect of the change of PDFs, which can alter the behaviour of other datasets that probe the large-x light quark and antiquark distributions. Full details about the trend in the fit quality for all datasets is given in App. D.

As far as the quality of the fit is concerned, we observe that for $\hat{Y} = 5 \cdot 10^{-5}$, the global fit is equivalent to the SM baseline, while as $\hat{Y}$ is increased to $15 \cdot 10^{-5}$ the quality of the fit deteriorates. This is due mostly to a worse description of the HL-LHC neutral current data (top left panel in Fig. 5.5) data, while the other datasets remain roughly equivalent. This is an indication that there is a bulk of data points in the global dataset that constrain the $\mathcal{L}^{\rm NC}$ luminosity behaviour at high-x and does not allow the PDF to shift and accommodate the HL-LHC Drell-Yan NC data. According to the selection criteria outlined in Sect. 5.1.3, the deterioration of both the χ^2 and the n_σ indicators would single out the high-mass Drell-Yan data and indicate that they are incompatible with the rest of the data included in the PDF fit. As a consequence, they would be excluded from the fit and no contamination would occur. Hence, in this scenario, $\hat{Y} = 5 \cdot 10^{-5}$ falls in the interval of NP values beyond which the disagreement in the data metrics would flag the incompatibility of the high-mass Drell-Yan tails with the rest of the datasets.

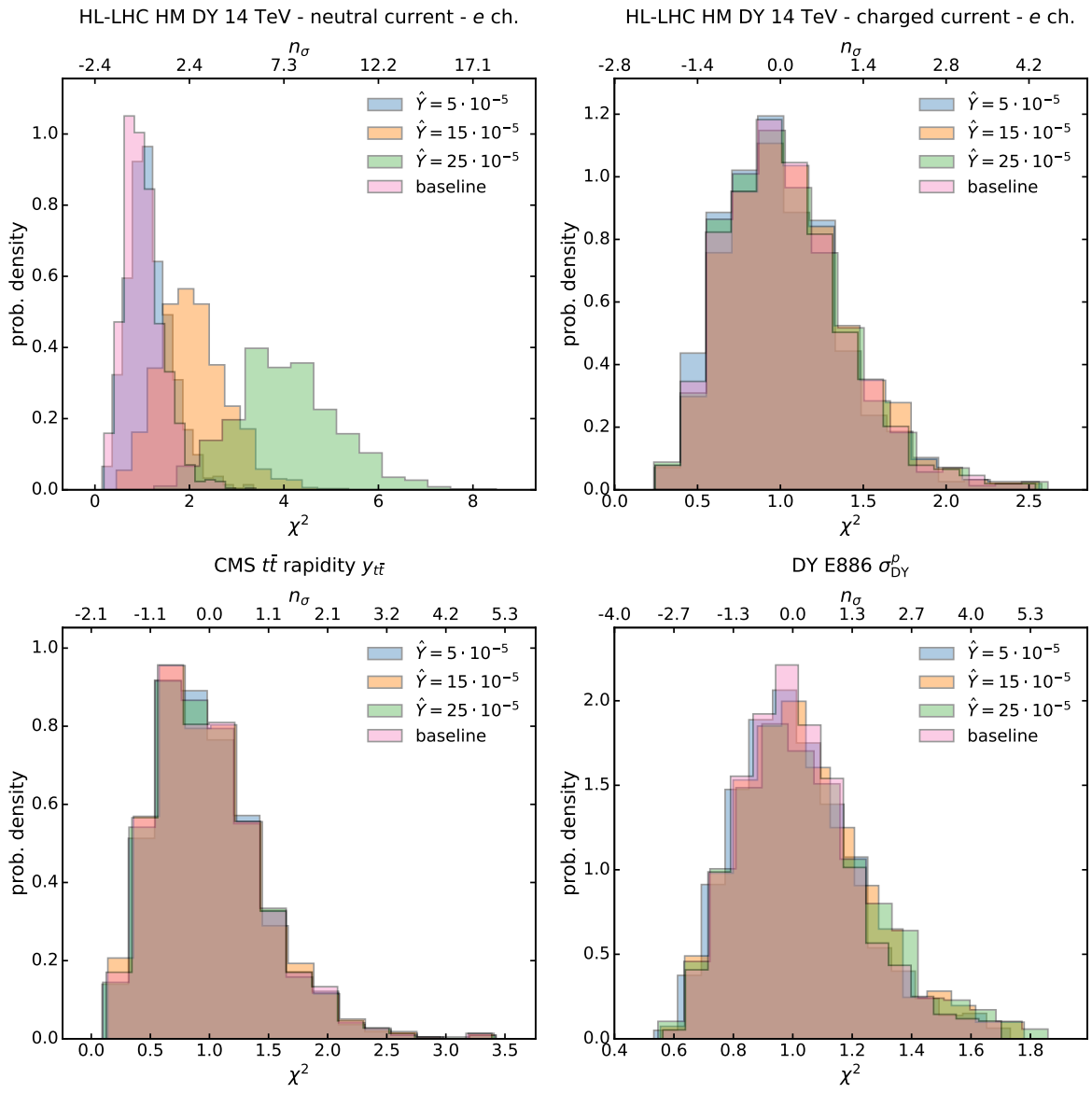


Figure 5.5: Distribution of χ^2 and n_σ for selected datasets in the $\hat{Y}$ contamination scenarios.

We now want to check whether, for such value, there is any significant shift in the relevant NC and CC parton luminosities at the HL-LHC centre-of-mass energy of $\sqrt{s} = 14$ TeV. They are displayed in Fig. 5.6.

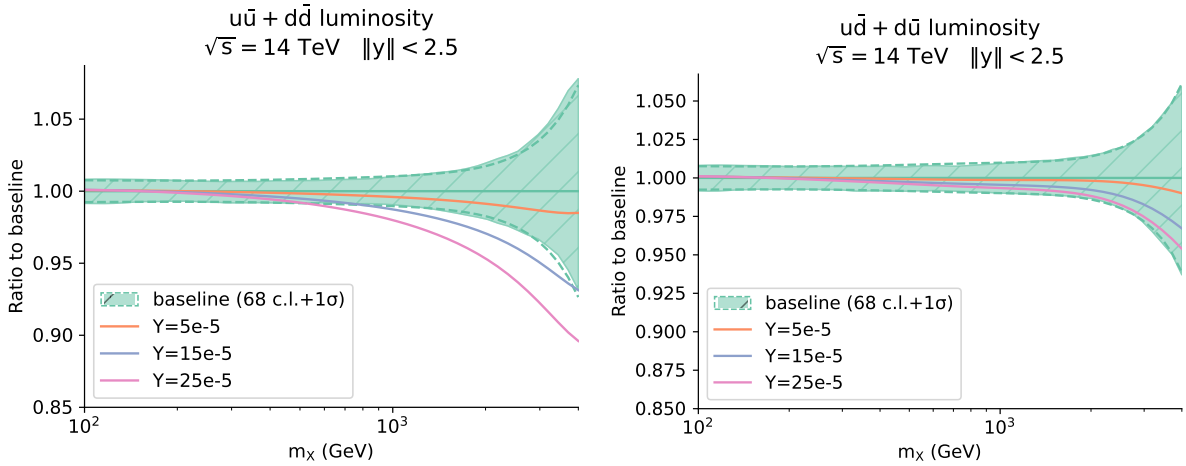


Figure 5.6: Contaminated versus baseline $\mathcal{L}^{\rm NC}$ and $\mathcal{L}^{\rm CC}$ (defined in Eq. (5.20), at $\sqrt{s} =$ 14 TeV in the central rapidity region. The results are normalised to the baseline SM luminosities and the 68% C.L. band is displayed. Contaminated PDFs have been obtained by fitting the MC data in which $\hat{Y} = 5 \cdot 10^{-5}$ (orange line), $\hat{Y} = 15 \cdot 10^{-5}$ (blue line) and $\hat{Y} = 25 \cdot 10^{-5}$ (pink line) has been injected.

We observe that in general the PDFs do not manage to shift much to accommodate the Z' induced contamination. The plots of the individual PDFs are displayed in App. E. In general the CC luminosity remains compatible with the baseline SM one up to large values of $\hat{Y}$, while, as soon as the NC luminosity manages to shift beyond the 1σ level, the fit quality of the NC high-mass data deteriorates. For the maximum value of new physics contamination that the PDFs can absorb in this scenario, $Y = 5 \cdot 10^{-5}$ (corresponding to a Z' mass above 30 TeV), the parton luminosity shift is contained within the baseline 1σ error bar. Overall, we see that there is a certain sturdiness in the fit, such that even in the presence of big $\hat{Y}$ values, the parton luminosity does not deviate much from the underlying law.

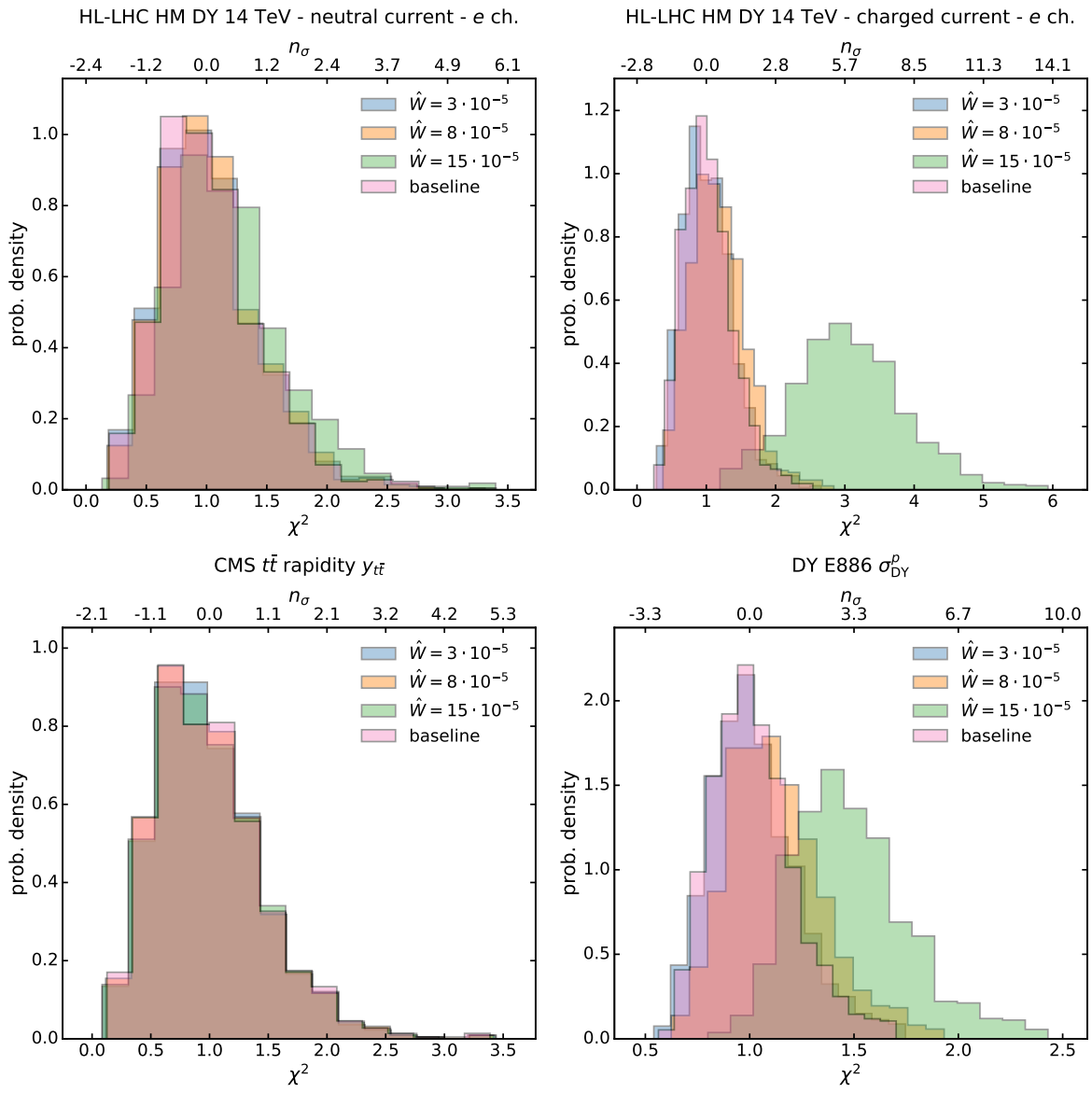


Figure 5.7: Distribution of χ^2 and n_σ for selected datasets in the $\hat{W}$ contamination scenarios.

(ii) Scenario II

In the flavour-universal W' model we inject three non-zero value of $\hat{W} = 3 \cdot 10^{-5}$, $8 \cdot 10^{-5}$, $15 \cdot 10^{-5}$. In Fig. 5.7 we display the $\chi^{2(k)}$ and $n_\sigma^{(k)}$ distributions across the 1000 random seeds k for a selection of the datasets included in each of the fits. In particular we display the datasets in which a shift occurs either because of the direct effect of the non-zero Wilson coefficients in the partonic cross sections (such as the high-mass Drell-Yan in the HL-LHC projections) or because of the indirect effect of the change of PDFs on other datasets that probe the large-x light quark and antiquark distributions. Full details about the trend in the fit quality for all datasets is given in App. D.

As far as the quality of the fit is concerned, we observe that up to $\hat{W} = 8 \cdot 10^{-5}$, the global fit shows equivalent behaviours to the SM baseline, while as $\hat{W}$ is increased to $15 \cdot 10^{-5}$, the quality of the fit markedly deteriorates. This is due mostly to a worse description of the HL-LHC charged current $e\nu_e$ (top right panel in Fig. 5.7) as well as the $\mu\nu_\mu$ data. It is interesting to observe that also the low-mass fixed-target Drell-Yan data from the E886 experiment experiences a deterioration in the fit quality due to the shift that occurs in the large-x quark and antiquark PDFs. Beyond $\hat{W} = 8 \cdot 10^{-5}$, according to the selection criteria outlined in Sect. 5.1.3, the deterioration of both the χ^2 and the n_σ indicators would point to the high-mass Drell-Yan data and indicate that they are incompatible with the bulk of the data included in the PDF fit and thus would be excluded from the fit and no contamination would occur. Hence, in this scenario, $\hat{W} = 8 \cdot 10^{-5}$ falls in the NP parameter region beyond which the disagreement in the data would unveil the presence of incompatibility of the high-mass Drell-Yan tails with the rest of the data.

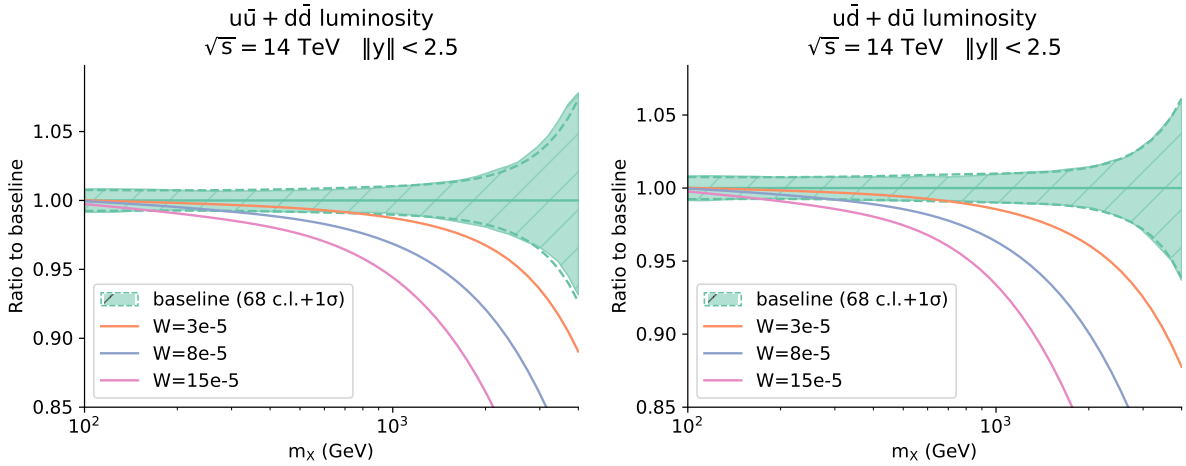


Figure 5.8: Same as Fig. 5.6 for $\hat{W} = 3 \cdot 10^{-5}$ (orange line), $\hat{W} = 8 \cdot 10^{-5}$ (blue line) and $\hat{W} = 15 \cdot 10^{-5}$ (pink line).

We now check whether, for such value, there is any significant shift in the PDFs and in the parton luminosities. Individual PDFs are displayed in App. E. In Fig. 5.8 we observe that in this scenario the NC and CC luminosities defined in Eq. (5.20) can both shift significantly in the high-mass region, even for low values of $\hat{W}$ ($m_{W'}$ above 20 TeV). Contrary to the case outlined in the $\hat{Y}$ scenario, the fit does have enough flexibility to absorb significant deviations in the high-mass Drell-Yan without impacting the rest of the dataset. In particular, until deviations become too large, the NC and CC sectors, which are both affected by the W' boson, manage to compensate each other.

(iii) Summary

Overall, we find that in Scenario I the presence of a new heavy Z' of about 18 TeV would affect the high-energy tails of the Drell-Yan distributions in such a way that they are no longer compatible with the bulk of the data included in a PDF analysis. On the other hand, in Scenario II, a model of new physics involving a W' of about 14 TeV would affect the high-energy tails of the Drell-Yan distributions in a way that can be compensated by the PDFs. As a result, if there is such a W' in Nature, then this would yield a good χ^2 for the high-mass Drell-Yan tails that one includes in a PDF fit as well as for the bulk of the data included in a PDF fit, but it would significantly modify PDFs. Thus, in this case new physics contamination does occur.

These results are in agreement with the results of Chapter 3, which generalise the analysis of Ref. [306] by allowing the PDFs to vary along with the $\hat{Y}$ and $\hat{W}$ coefficients, finding less stringent constraints from the same HL-LHC projections. In particular, it was found that $\hat{W} = 8 \cdot 10^{-5}$ would have been excluded by the HL-LHC under the assumption of SM PDFs, but that this value of $\hat{W}$ was allowed by the constraints at 95% CL obtained by varying the PDFs along with the SMEFT. Chapter 3 also indicated that the impact of varying the PDFs along with the $\hat{W}$ coefficient was more significant than the impact in the $\hat{Y}$ direction, indicating a higher possibility to absorb the effects of new physics into the PDFs in the $\hat{W}$ direction.

Comparing the two scenarios considered in this section, one might wonder why the Z' scenario does not yield any contamination, while the W' does. Looking at the effect of the Z' and W' bosons on the observables included in a PDF fit (see Eqs. (5.9) and (5.15) respectively), we see that the main difference lies in the fact that the Z' scenario only affects the NC DY high-mass data, while the W' scenario affects both the NC and the CC DY high-mass data. Hence, in the former scenario, the shift required in $\mathcal{L}^{\rm NC} \equiv (u\bar{u} + d\bar{d})$ to accommodate the effect of a Z' in the tail of the m_{ll} distribution would cause a shift in $\mathcal{L}^{\rm CC} \equiv (u\bar{d} + d\bar{u})$, thus spoiling its agreement with the data, in particular the tails of the m_T distribution – which is unaffected by the presence of a Z'.

On the other hand, in the W' scenario, the shift in the $(u\bar{u} + d\bar{d})$ parton channel that accommodates the effect of a W' in the tail of the NC DY m_{ll} distribution is compensated by the shift in the $(u\bar{d} + d\bar{u})$ parton channel that accommodates the presence of a W' in the tail of the CC DY m_T distribution (as, in this scenario, they are both affected by new physics). It is as if there is a flat direction in the luminosity versus the matrix element space. This continues until, for sufficiently large $\hat{W}$, a critical point is reached in which the two effects do not manage to compensate each other as they start affecting significantly the luminosities at lower $\tau = M/\sqrt{s}$, hence spoiling the agreement with the other less precise datasets included in a PDF fit which are sensitive to large-x anti-quarks.

To see this more clearly, we plot in Fig. 5.9 the data-theory comparison for the HL-LHC

NC and CC Drell-Yan Monte Carlo data that we include in the fit.

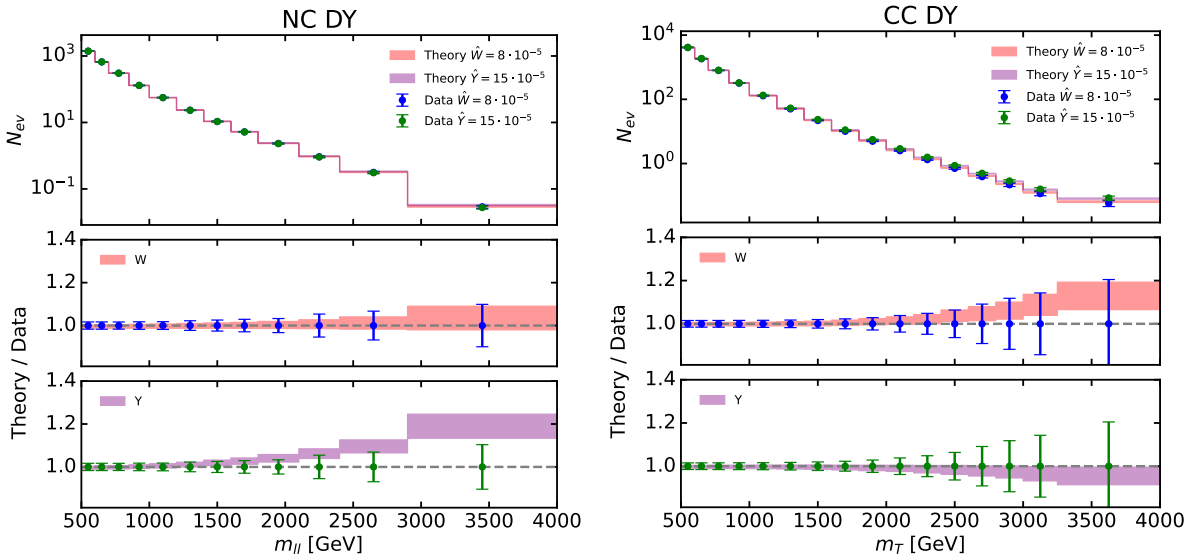


Figure 5.9: For two of the representative scenarios that we consider, $\hat{W} = 8 \cdot 10^{-5}$ and $\hat{Y} = 15 \cdot 10^{-5}$, we show the comparison between the data expected in presence of new physics ('data' points) and the SM theory predictions obtained with the potentially contaminated PDFs ('Theory' bands). Left panel: NC Drell-Yan m_{ll} distribution. Right panel: CC Drell-Yan m_T distribution.

The points labelled as 'data' correspond to the 'truth' in presence of the new physics, namely they are obtained by convoluting the DY prediction with non-zero $\hat{Y}, \hat{W}$ parameters with a non-contaminated PDF set. The bands labelled as 'theory' represent the theoretical predictions for pure SM DY production, but obtained with the PDFs fitted with the inclusion of the DY data modified by the effect of non-zero $\hat{Y}, \hat{W}$ parameters. We observe that the SM predictions obtained with the contaminated PDFs do fit the data well in the case of $\hat{W} = 8 \cdot 10^{-5}$, because the significant depletion of the $(u\bar{u} + d\bar{d})$ and $(u\bar{d} + d\bar{u})$ parton luminosities observed in Fig. 5.8 compensates the enhancement in the partonic cross section observed in Fig. 5.2. This is not the case for $\hat{Y} = 15 \cdot 10^{-5}$, where instead the much milder modification of the parton luminosities observed in Fig. 5.6 does not manage to compensate the enhancement of the partonic cross section observed in Fig. 5.1. We can also notice that $\hat{W} = 8 \cdot 10^{-5}$ is within a region in the $\hat{W}$ parameter space beyond which the parton luminosities do not manage to move enough to compensate the shift in the matrix elements of the m_T distribution. To find the exact critical value of $\hat{W}$ one would need a finer scan. Analogously, $\hat{Y} = 15 \cdot 10^{-5}$ is in the region of $\hat{Y}$ such that contamination in the PDFs does not occur. However, these values have been determined assuming a given statistical uncertainty in the distributions; the regions in which these values fall clearly depends on the actual statistical uncertainty that the m_T and m_{ll} distributions will reach in the HL-LHC phase.

5.3.3 Consequence of new physics contamination in PDF fits

In the previous section, we showed that in the presence of heavy new physics effects in DY observables, the flexible PDF parametrisation is able to accommodate the deviations and absorb the effects coming from the new interactions. In particular, we observe that when data are contaminated with the presence of a W', we generally find good fits and are able to accommodate even large deviations from the SM. It is however worth reminding that the leading source of contaminated data are the HL-LHC projection, as present data would not be as susceptible to the W' effects. Hence, from now on we will focus on the scenario in which data include the presence of a heavy W' which induces a modified interaction parametrised by the $\hat{W}$ parameter with value $\hat{W} = 8 \cdot 10^{-5}$.

In this section we examine the consequences of using unknowingly contaminated PDFs, and the implications of this for possible new physics searches. The first interesting consequence is that, if we use the contaminated PDF as an input set in a SMEFT study of HL-LHC projected data to gather knowledge on the $\hat{W}$ parameter, we find that the analysis excludes the 'true' value of the SMEFT coefficients that the data should reveal. Indeed, in Fig. 5.10 we observe that, in both scenarios under consideration, and in particular for the one corresponding to $\hat{W} = 8 \cdot 10^{-5}$, the 95% C.L. bounds on the Wilson coefficients that one would extract from the precise HL-LHC data discussed in Sect. 5.3.2 would agree with the SM and would not contain the true "values" of the underlying law of Nature that the data should reveal. In fact, the measured value would exclude the true value with a significance that ranges from $\sim 1.5\,\sigma$ to $\sim 4.5\,\sigma$. A comparison of whether the bounds generated by the different contaminated PDFs considered in this study contain the true value is shown in Fig. 5.10.

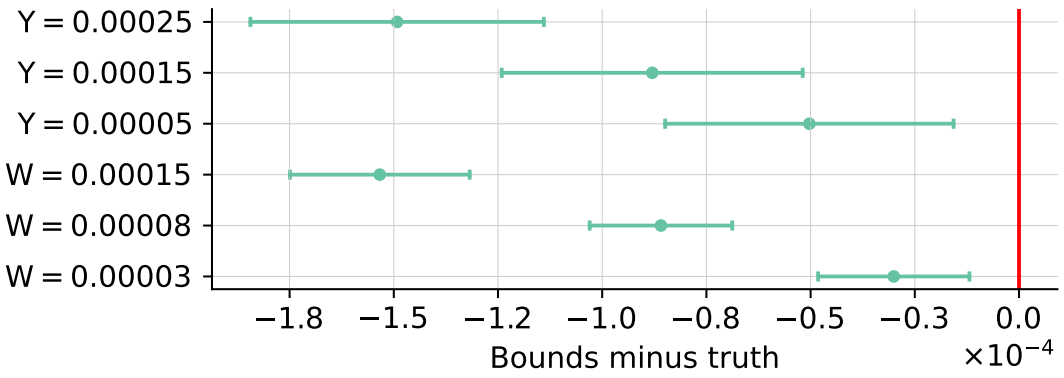


Figure 5.10: A comparison of the 68% bounds obtained using different contaminated PDFs to fit the $\hat{W}$, $\hat{Y}$ parameters to HL-LHC high-mass Drell-Yan projected data, relative to the true values of $\hat{W}, \hat{Y}$. In some cases, the true value is not contained in the 95% confidence level bounds.

This is all very expected, as the quark-antiquark luminosity for this specific scenario

does exhibit signs of new physics absorption in a significant amount, as can be seen in Fig. 5.8. As a matter of fact, we expect all data that entered in the PDF fit to be well described by the combination of the PDF set and the SM theory. This simple fact is once again reminding us that it might be dangerous to perform SMEFT studies on overlapping datasets and that simultaneous studies should be preferred or, at least, a conservative approach with disjoint datasets should be undertaken. This was for instance discussed in Ref. [103], where it was shown that by means of a simultaneous study, one is able to recover both the underlying true PDFs and the presence of a new interaction. It is worth mentioning that the use of conservative PDF sets, while appealing given the simplicity, might also come with its own shortcomings, see Chapter 3 and Ref. [40] for detailed studies on the matter in the Drell-Yan sector and the top quark sector respectively. In particular, the extrapolated PDFs might both underestimate the error band and have a significant bias.

We now turn to study what would be the effects of the contaminated PDFs in observables and processes that did not enter the PDF fit. We focus in particular in the EW sector, given its relevance for NP searches and the fact that the contaminated PDFs show deviations from the true PDFs mostly in the quark-antiquark luminosities, which are particularly relevant for theoretical predictions involving EW interactions. The study is performed by producing simulated data according to the true laws of Nature, i.e. the true PDFs of choice and the SM + $\hat{W} = 8 \cdot 10^{-5}$ in the matrix elements.

In particular, we produce MC data for several diboson processes, including H production in association with EW bosons. Given that the $\hat{W}$ operator induces only four-fermion interactions, $\hat{W}$ does not have an effect on these observables, and the hard scattering amplitudes are given by the SM ones. For each observable we build HL-LHC projections and devise bins with the objective of probing the high-energy tails of the distributions, scouting for new physics effects that we although know do not exist in the "true" law of Nature. We then produce predictions by convoluting the contaminated PDF set obtained with a value of $\hat{W} = 8 \cdot 10^{-5}$ and the SM matrix elements. Given our knowledge of the "true" law of Nature, the possible deviations between theory and data are therefore only a consequence of the shift in the PDFs coming from the contaminated Drell-Yan data. Whenever in the presence of W bosons, we decided to split the contributions of W^+X and W^-X as they probe different luminosities and in particular, from the contaminated fits, we know that the luminosity $u\bar{d}$ is deviating more severely than $d\bar{u}$ from the true luminosity.

Both SM theory and data have been produced at NLO in QCD making use of the Monte Carlo generator `MadGraph5_aMC@NLO`. In the case of ZH production, the gluon fusion channel has also been taken into account. Data are obtained by fluctuating around the central value, assuming a Gaussian distribution with total covariance matrix given

by the sum of the statistical, luminosity and systematic covariance matrices. Regarding the theory predictions, we also provide an estimate of the PDF uncertainty. We assume a luminosity of 3 ab^{-1} and we estimate the systematic uncertainties on each observable by referring to the experimental papers [312, 313, 314] . These systematic uncertainties can be experimental or come from other additional sources such as background and signal theoretical calculations. We also include estimates of the luminosity uncertainty by taking as a reference the CMS measurement at 13 TeV [104]. Statistical uncertainties are given by $\sqrt{N}$, where N is the number of expected events in each bin. Performing a fully realistic simulation, with acceptance cuts and detector effects, is beyond the scope of the current study, and we simply simulate events at parton level and apply the branching ratios into relevant decay channels. Specifically, in the case of W bosons we apply a $Br(W \to l\nu) = 0.213$ with $l = e$, μ, for the H boson we consider $Br(H \to b\bar{b}) = 0.582$ and for the Z boson we have $Br(Z \to l^+l^-) = 0.066$ with $l = e$, μ [307]. To combine multiple sources of uncertainty we add them in quadrature.

	HL-LHC		Stat. improved	
Dataset	$\chi^2/n_{\rm dat}$	n_σ	$\chi^2/n_{\rm dat}$	n_σ
W^+H	1.17	0.41	1.77	1.97
W^-H	1.08	0.19	1.08	0.19
W^+Z	1.08	0.19	1.49	1.20
W^-Z	0.99	-0.03	1.02	0.05
ZH	1.19	0.44	1.67	1.58
W^+W^-	2.19	3.04	2.69	4.31
VBF $\to$ H	0.70	-0.74	0.62	-0.90

Table 5.4: Values of the χ^2 and n_σ for the projected observables at HL-LHC in the EW sector. In the left column we report the values from a realistic estimate of the statistical uncertainties, while in the right columns we show what would be obtained if statistics were to improve by a factor 10.

In Table 5.4, for each process considered, we collect the computed χ^2 and the corresponding value of n_σ. These numbers are obtained by performing several fluctuations of the data and then taking the average χ^2 from all the replicas. As a consequence, the quoted χ^2 are considered the expected χ^2 and are not associated to a specific random fluctuation. The numbers are provided both in a realistic scenario, with a reasonable estimate of the statistical uncertainties, and in a scenario in which the statistics is improved by a factor 10. The latter could be both the result of an increased luminosity and/or additional decay channels of the EW bosons, e.g. decays into jets. As it can be seen by inspection of the table, the processes which would lead to the most notable deviations between data and theory are W^+H and W^+W^-, with the latter being in significant tension already in the scenario of a realistic uncertainty estimation. With improved statistics, slight tensions

start to appear in ZH and W^+Z, both exhibiting a deviation just above 1σ. Interestingly, the clear smoking gun process here seems to be W^+W^-, which just by itself would point towards a significant tension with the SM, which could potentially and erroneously be interpreted in terms of new interactions.

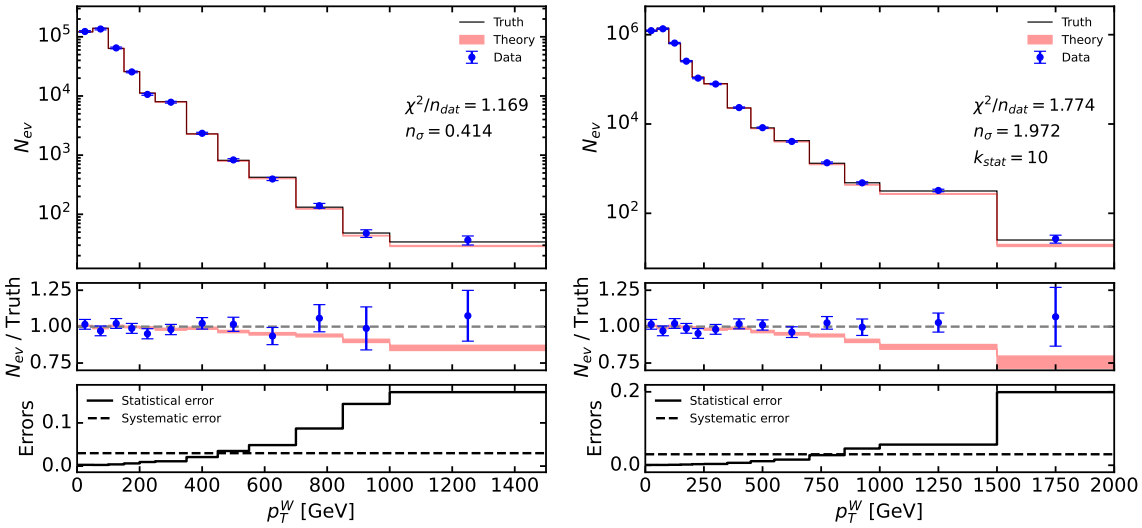


Figure 5.11: Predictions (with contaminated PDF) for W^+H at the HL-LHC compared with the projected data. Left: HL-LHC projection. Right: statistics improved by a factor 10 (futuristic scenario). In the latter, an additional bin is added at high energy to take advantage of the additional expected events.

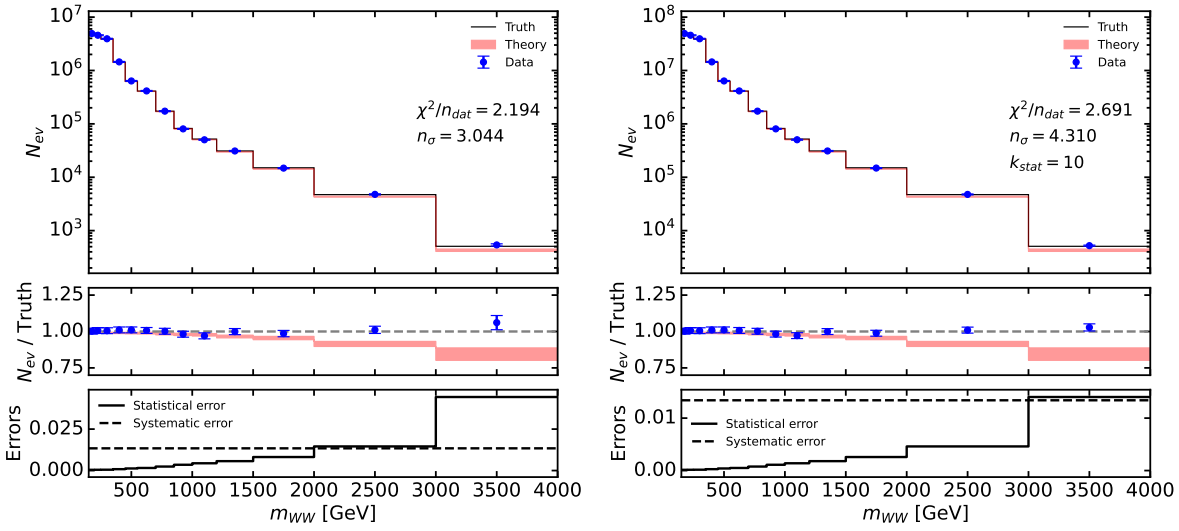


Figure 5.12: Predictions (with contaminated PDF) for W^+W^- at the HL-LHC compared with the projected data. Left: HL-LHC projection. Right: statistics improved by a factor 10 (futuristic scenario).

In Fig. 5.11 and Fig. 5.12 we show plots of the two most affected observables considered in this section, namely W^+H and W^+W^- respectively. While in all other processes the

deviations between the true central values and the theoretical predictions obtained with the contaminated PDFs are limited, in the case of W^+H and W^+W^-, they are substantial. It is clear that the true limiting factor is that as soon as we are in the high energy tails of the distributions and potentially sensitive to the PDF contamination, the pseudo data become statistically dominated and therefore we lose resolution. This is particularly true in the case of W^+H, while W^+W^- is predicted to have an higher number of events and could potentially probe higher energies.

We also assess the ratios W^+Z/W^-Z and W^+H/W^-H, and observe that in this case the deviations resulting from contaminated PDFs are no longer visible when taking these ratios. In general the ratios cancel the effect of any possible contamination in the parton luminosities if they are correlated. The fact that the effect disappears is a proof that the $u\bar{d}$ and $d\bar{u}$ luminosities are highly correlated and the contamination effects are compatible.

In summary, the PDF contamination has the potential to generate substantial deviations in observables and processes generally considered to be good portals to new physics, which could nonetheless be unaffected by the presence of heavy states at the current probed energies, as in the scenarios considered in this work.

5.4 How to disentangle New Physics effects

In this section we discuss several strategies that can be devised in order to disentangle New Physics effects in a global fit of PDFs. In Sect. 5.4.1 we start by assessing the potential of precise on-shell forward vector boson production data in the HL-LHC phase and check whether their inclusion in a PDF fit helps disentangling New Physics effects in the high-mass Drell-Yan tails.

We then turn to analyse the behaviour of suitable observable ratios in Sect. 5.4.2 and we will see that such ratios will indicate the presence of New Physics in the observables that are affected by it, although they would not be able to distinguish between the two observables that enter the ratio. Finally in Sect. 5.4.3 we will determine the observables in current PDF fits that are correlated to the large-x antiquarks and we will highlight the signs of tension with the 'contaminated' high-mass Drell-Yan data via suitably devised weighted fits. The result of these tests points to the need for the inclusion of independent low-energy/large-x constraints in future PDF analyses, if one wants to safely exploit the constraining power of high-energy data without inadvertently absorbing signs of New Physics in the high-energy tails.

5.4.1 On-shell forward boson production

The most obvious way to disentangle any possible contamination effects in the PDF is the inclusion of observables that probe the large-x region in the PDFs at low energies, where

NP-induced energy growing effects are not present. In this section we assess whether the inclusion of precise forward LHCb distributions measured at the W and Z on-shell energy at HL-LHC might help spotting NP-induced inconsistencies in the high-mass distributions measured by ATLAS and CMS.

In order to test this, we compute HL-LHC projections for LHCb, taking 0.3 ab^{-1} as benchmark luminosity [292] and focusing on the forward production of W/Z. The Z boson is produced on-shell (60 GeV $< m_{ll} <$ 120 GeV), while no explicit cuts are applied on the transverse mass m_T in the case of W boson production decaying into a muon and a muonic neutrino, which is dominated by the mass-shell region. We impose the LHCb forward cuts on the lepton transverse momentum ($p_T^l > 20$ GeV) and on both the Z rapidity and pseudo-rapidity of the μ originated by W ($2.0 < |y_{Z,\mu}| < 4.5$). Fig. 5.13 shows a comparison between the pseudo-data generated with the "true" PDFs and NP-corrected matrix elements,[1] and the theory predictions obtained with the $\hat{W} = 8 \cdot 10^{-5}$ contaminated fit and the SM matrix element, for each of the two processes. We observe that there are no significant deviations between the theory predictions obtained from a contaminated PDF set and the true underlying law. Intuitively this can be understood, as the produced

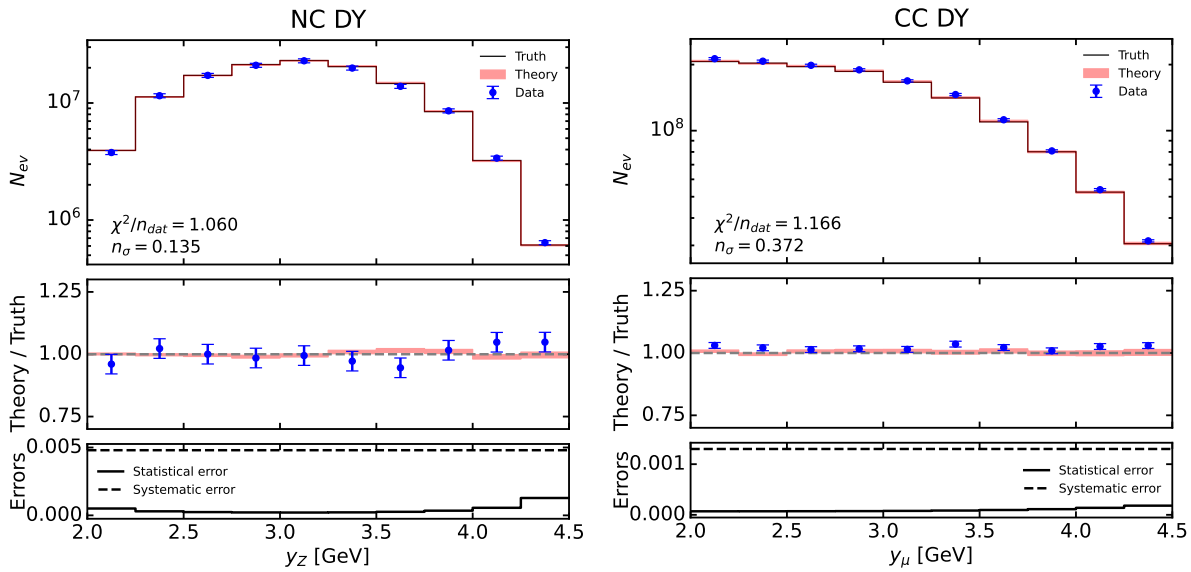


Figure 5.13: Predictions (with $\hat{W} = 8 \cdot 10^{-5}$ contaminated PDF) for forward vector bosons production in the HL-LHC phase at LHCb compared with the projected data. Left panel: on-shell Z production cross section as a function of the Z boson rapidity y_Z. Right panel: W production cross section as a function of the final-state muon pseudo-rapidity y_μ.

leptons are in the forward region measured at LHCb, and one of the initial partons must have more longitudinal momentum than the other.

To visualise more precisely the regions in x that are constrained by a measurement of

[1]Note that at the energy probed by the forward W/Z production the NP contribution associated to the presence of a W' boson is negligible

a given final state at the energy $E \sim m_X$ and at a given rapidity y, we display the scatter plot for $x_{1,2} = m_X/\sqrt{s}\,\exp(\pm y)$ in the large-m_X and central region, namely $|y| < 2.0$ and $1\,\text{TeV} < m_X < 4\,\text{TeV}$, and compare it to the low-to-intermediate-m_X and forward rapidity region, namely $2.0 < |y| < 4.5$ and $10\,\text{GeV} < m_X < 1\,\text{TeV}$.

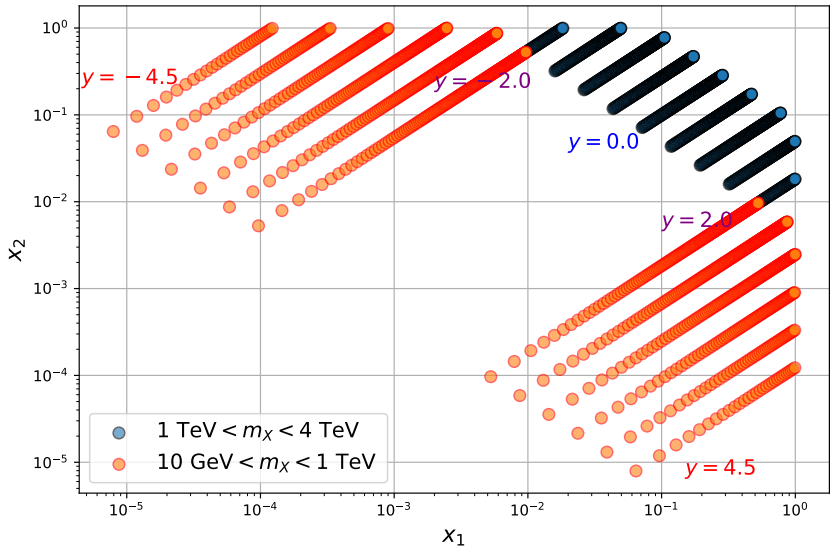


Figure 5.14: Leading order kinematic plot of $x_{1,2} = m_X/\sqrt{s}\,\exp(\pm y)$ in the large-M and central region, $|y| < 2.0$, $1\,\text{TeV} < m_X < 4\,\text{TeV}$, (blue bots and in the low-intermediate-m_X and forward region, $2.0 < |y| < 4.5$ $10\,\text{GeV} < m_X < 1\,\text{TeV}$ (red dots) . Here $\sqrt{s} = 14$ TeV.

We can see that, while the measurements of large-invariant mass objects in the central rapidity region constrain solely the large-x region, and where both partons carry a fraction x of proton's momentum in the $0.01 \lesssim x \lesssim 0.8$, the low-to-intermediate invariant mass region in the forward rapidity region constrains both the small and the large x region, given that at $|y| \approx 4.0$ the x-region probed is around $0.1 \lesssim x \lesssim 0.8$ for one parton and around $10^{-5} \lesssim x \lesssim 10^{-4}$ for the other parton. Given that the valence quarks are much more abundant at large x than the sea quarks, in most collisions the up or down quark will be the partons carrying a large fraction x of the proton's momentum, while the antiquarks will carry a small fraction x. Hence, this observable will not be sensitive to the shift in the large-x anti-up and anti-down that the global PDF fit yields in order to compensate the effect of NP in the tails.

5.4.2 Observable ratio

In order to disentangle PDF contamination, another quantity worth studying is the ratio between observables whose processes have similar parton channels. Indeed, in this case the impact of the PDF is much reduced and any discrepancy between theory and data predictions can be more confidently attributed to new physics in the partonic cross-section. Practically, a deviation would mean that one of the two datasets involved in the ratio is 'contaminated' by new physics and should therefore be excluded from the PDF fit.

We have studied the ratio between the number of events in WW production and Neutral Current Drell-Yan (NC DY), as well as between WH production and Charged Current Drell-Yan (CC DY). In each pair both processes are initiated from the same parton channels.

The Drell-Yan events we use are displayed in Fig. 5.9. The diboson events can be seen in Fig. 5.12 for WW and in Fig. 5.11 for W^+H. However, note that we also include the W^-H channel to measure the ratio of WH and CC DY here. We plot the ratio of those quantities in Fig. 5.15. We compare theory and data predictions where, as in Fig. 5.9, data corresponds to a baseline PDF and a BSM partonic cross-section ($f_{\text{Baseline}} \otimes \hat{\sigma}_{BSM}$) and theory is computed from a contaminated PDF and a SM partonic cross-section ($f_{\text{Cont}} \otimes \hat{\sigma}_{SM}$). We also compare those results to K-factors which are obtained by taking the ratio of Drell-Yan BSM predictions over the SM ones. Practically the K-factors are a ratio of their respective partonic cross-section ($K = \hat{\sigma}^{DY}_{BSM}/\hat{\sigma}^{DY}_{SM}$).

We see in both cases a deviation between theory and data predictions growing with the energy. The uncertainties are smaller in the ratio WW/NC DY which allows the discrepancy to be over 1σ in the last bin. Furthermore, we also witness that the deviation follows the K-factors which reinforces our initial assumption that using ratio greatly diminishes the impact of the PDFs. As we mentioned earlier, the lesson we can get from this plot is that there is some new physics in either the DY or the diboson datasets. Unfortunately, without further information it is not possible to identify in which of those datasets the new physics is. Therefore, with just this plot in hand, the only reasonable decision would be to exclude both the datasets involved in the ratio where the deviation is observed from the fit. The downside of this disentangling method is that it might worsen the overall quality of the fit and increase the PDF uncertainties in certain regions of the parameter space. However, it proves to be an efficient solution against the sort of contamination we studied. Indeed, by excluding the DY datasets in this case, one would exclude the contamination we manually introduced there from the PDF fit.

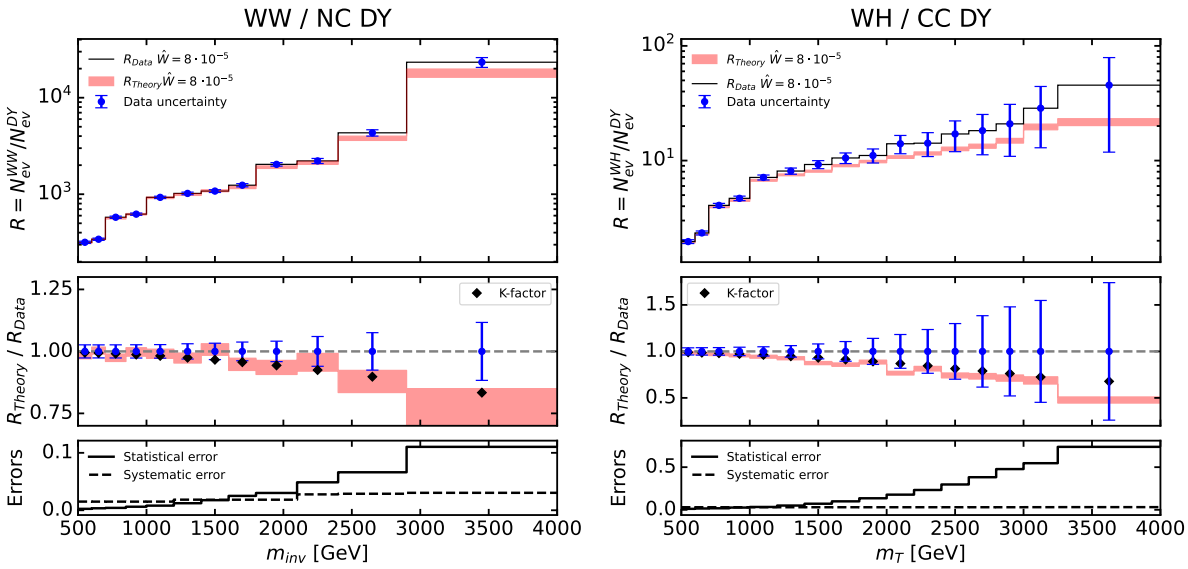


Figure 5.15: Ratio between diboson production and Drell-Yan processes for the HL-LHC predictions. On the left, ratio of W^+W^- to NC DY binned in invariant mass and on the right, ratio of WH to CC DY binned in transverse mass. In the top panel we plot the ratios of number of events for data and theory predictions. In the middle panel, we plot the ratio of those ratios (theory over data) alongside the K-factors. The lower panel displays the uncertainties.

5.4.3 Alternative constraints on large-x anti-quarks

In Sect. 5.4.1, it was shown that the inclusion of precise on-shell forward W and Z production measurements does not disentangle the contamination that New Physics in the high-energy tails might yield. In this section, we ask ourselves whether there are any other future low-energy observable that might constrain large-x antiquarks and show tension with the high-energy data in case the latter are affected by NP-induced incompatibilities.

We start by looking at the correlation between the data that are currently included in our baseline PDF fit and the various PDF flavours. To assess the level of correlation, we plot the correlation defined in Ref. [315]. The correlation function is defined as

$$\rho(j,x,\mathcal{O}) \equiv \frac{N_{\rm rep}}{N_{\rm rep}-1}\left(\frac{\langle f_j(x,Q)\,\mathcal{O}\rangle_{\rm reps}-\langle f_j(x,Q)\rangle_{\rm reps}\langle\mathcal{O}\rangle_{\rm reps}}{\Delta_{\rm PDF}f(x,Q)\,\Delta_{\rm PDF}\mathcal{O}}\right), \tag{5.22}$$

where the PDFs are evaluated at a given scale Q and the observable $\mathcal{O}$ is computed with the set of PDFs f, j is the PDF flavour, $N_{\rm rep}$ is the number of replicas in the baseline PDF set and $\Delta_{\rm PDF}$ are the PDF uncertainties. In Figs. 5.16 and 5.17 we show the correlation between the PDFs in the flavour basis and the observables which are strongly correlated with the anti-quark distributions. The region highlighted in blue is the region in x such that the correlation coefficient defined in Eq. (5.22) is larger than $0.9\,\rho_{\rm max}$, where $\rho_{\rm max}$

is the maximum value that the correlation coefficient takes over the grid of points in x and over the flavours j. From Fig. 5.16 we observe that while the largest invariant mass bins of the HL-LHC NC are most strongly correlated with the up anti-quark distribution in the $10^{-2} \lesssim x \lesssim 3 \cdot 10^{-1}$ region, the HL-LHC CC, particularly the lowest invariant mass bins, are most strongly correlated with the down anti-quark distribution in the $7 \cdot 10^{-3} \lesssim x \lesssim 5 \cdot 10^{-2}$ region. This observation is quite interesting as it gives us a further insight on the difference between the Z' and the W' scenarios discussed at the end of Sect. 4.2. Indeed the W' scenario affecting both the NC and CC distributions manages to compensate the $\bar{u}$ shift with the $\bar{d}$ shift in a slightly smaller region, hence the successful contamination.

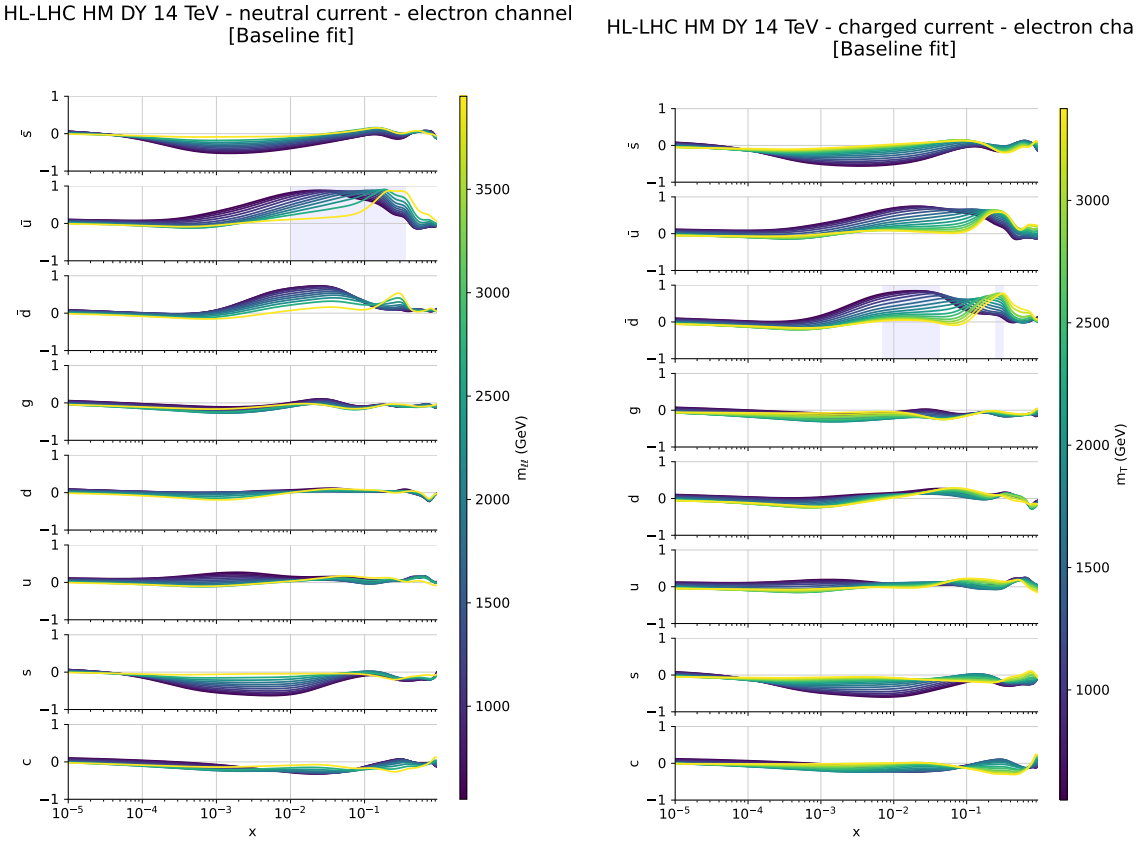


Figure 5.16: Correlation coefficient ρ defined in Eq. (5.22) between the flavour PDFs of the baseline set and the HL-LHC neutral current Drell-Yan data (left panel); the HL-LHC charged current Drell-Yan data (right panel). The highlighted region corresponds to $\rho > 0.9\,\rho_{\rm max}$.

We now ask ourselves whether there are other observables that display a similar correlation pattern with the light anti-quark distributions. In Fig. 5.17 we show the three most interesting showcases. In the left panel, we see that that the FNAL E866/NuSea measurements of the Drell-Yan muon pair production cross section from an 800 GeV proton beam incident on proton and deuterium targets [316] yields constraints on the the ratio of

anti-down to anti-up quark distributions in the proton in the large Bjorken-x region and the correlation is particularly strong with the anti-up in the $5 \cdot 10^{-2} \lesssim x \lesssim 3 \cdot 10^{-1}$ region. On the central panel we see that the Tevatron D0 muon charge asymmetry [317] exhibits a strong correlation with the up anti-quark around $x \approx 0.3$ and the down quark around $x \approx 0.1$. This is understood, as by charge conjugation the anti-up distribution of the proton corresponds to the up distribution of the anti-proton. Finally, on the right panel we see that the precise ATLAS measurements of the W and Z differential cross-section at $\sqrt{s} = 7$ TeV [318] have a strong constraining power on the up anti-quark in a slightly lower x region around $3 \cdot 10^{-3} \lesssim x \lesssim 2 \cdot 10^{-2}$.

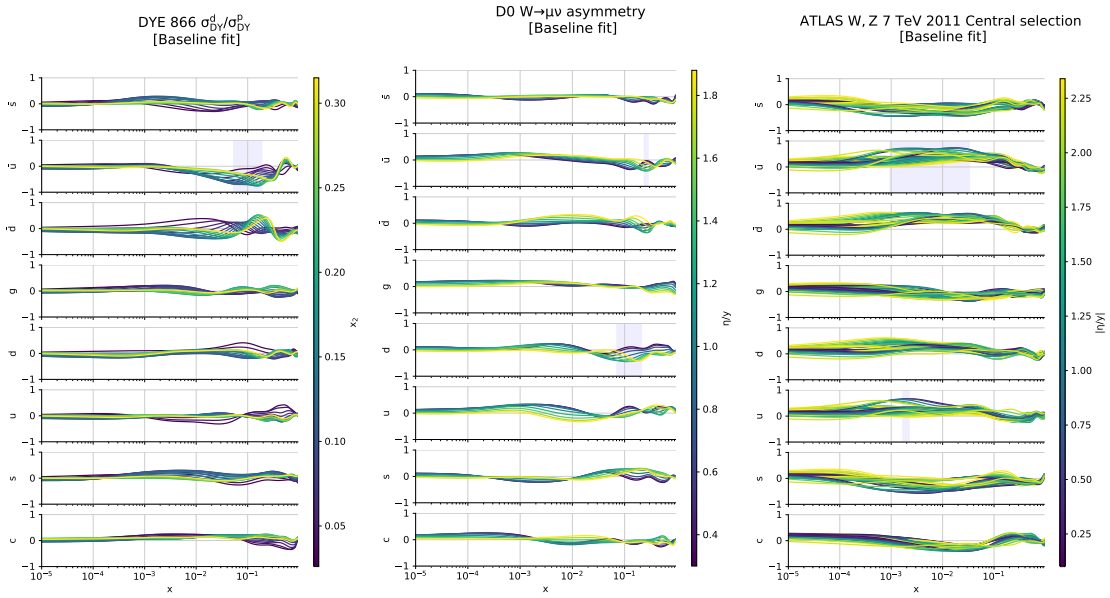


Figure 5.17: Same as Fig. 5.16 for the FNAL E866 data measuring the ratio between low energy Drell-Yan muon pair production on proton and deuteron targets [316] (left panel), the Tevatron D0 muon charge asymmetry [317] (central panel) and the ATLAS measurements of the W and Z differential cross-section at $\sqrt{s} = 7$ TeV in the central rapidity region[318] (right panel).

The results presented in Sect. 5.3 show that the tension with the low-energy datasets that constrain the same region in x as the high-mass Drell-Yan HL-LHC data is not strong enough to flag the HL-LHC datasets. Hence, the conditions highlighted in Sect. 5.1.3 are necessary in order to determine a bulk of maximally consistent datasets, but they are not sufficient, as they still allow New Physics contamination to go undetected. A way to emphasise the tension is to produce *weighted fits* which give a larger weight to the high-energy data that are affected by New Physics effects. The rationale behind this is that, if some energy-growing effect associated to the presence of New Physics in the data shows up in the tails of the distributions, PDFs might accommodate this effect without deteriorating the agreement with the other datasets only up to a point. If the datasets that

are affected by New Physics are given a large weight in the fit, the tension with the datasets constraining the large-x region that are perfectly described by the SM could in principle get worse. Hence by giving a larger weight to a specific NP-affected dataset, the disagreement with the datasets highlighted above should become more apparent. Depending on the kind of New Physics model that Nature might display in the data, the effect might affect either of the three classes of the high energy processes entering PDF fits, namely: (i) jets and dijets, (ii) top and (iii) Drell-Yan. In our example, in order to emphasise the tension with the low-energy Drell-Yan data and the Tevatron data, we would have to give more weight to the HL-LHC high-mass Drell Yan data. However we observe that, although the χ^2 of the HL-LHC Drell-Yan further improves and the ones of the highlighted data deteriorates, the level of deterioration is never strong enough to flag the tension. The result of this test points to the fact that one should include independent and more precise low-energy/large-x constraints in future PDF analyses, if one wants to safely exploit the constraining power of high-energy data without inadvertently absorbing signs of New Physics in the high-energy tails. In this sense the low-energy EIC programme, as well as other low-energy data which are not exploited in the standard PDF global fits, such as JLAB or SeaQuest data, will be a precious input in future PDF analyses, alongside the constraints from lattice measurements.

Chapter 6

The Monte Carlo replica method in global fits

Yesterday,
All my troubles seemed so far away,
Now it looks as if they're here to stay.

---

from Yesterday,
by The Beatles

[This chapter is based on an upcoming publication, worked on in collaboration with Maeve Madigan and Luca Mantani. The calculation presented in Sect. 6.2 is my own original work.]

Throughout this thesis, particularly in Chapters 3 and 4, we have used the Monte Carlo replica method for propagation of experimental errors onto the PDFs and theory parameters in our fits. However, as described in Sect. 4.7, it is possible to demonstrate that this method does not faithfully propagate errors in the case where non-linear terms dominate in our theory predictions.

In this chapter, we present a more complete description of the phenomenon compared to Sect. 4.7, generalising fully to an arbitrary number of (possibly correlated) data points, theory parameters and an arbitrary (smooth) theory function $\mathbf{t}$. We begin by stating the expected Bayesian results in the multivariable case in Sect. 6.1. Subsequently, in Sect. 6.2, we present a detailed calculation of the posterior distribution that results from an application of the Monte Carlo replica method, and compare with the Bayesian result of Sect. 6.1. Finally, in Sect. 6.3, we discuss the implications this result could have for various global fits, including recent PDF fits, and we propose a course of action for the future.

6.1 Bayesian interval estimation in the multivariable case

In Sect. 4.7, we introduced error propagation using Bayesian methods for a single data point and a single theory parameter, with a quadratic theory prediction. In this brief section, we state the results in the more general context of multiple (possibly correlated) data points and multiple theory parameters, with a general (smooth) theory prediction.

Let us assume that we are fitting a vector of theory parameters $\mathbf{c} \in \mathbb{R}^{N_{\text{param}}}$ (which now may comprise PDF parameters alongside EFT parameters, for example - in Sect. 4.7 we worked exclusively with EFT coefficients), given data $\mathbf{d} \in \mathbb{R}^{N_{\text{dat}}}$. The theory predictions are a general (smooth[1]) function $\mathbf{t} : \mathbb{R}^{N_{\text{param}}} \to \mathbb{R}^{N_{\text{dat}}}$, and we assume that the data is normally distributed according to:

$$\mathbf{d} \sim \mathcal{N}(\mathbf{t}(\mathbf{c}), \Sigma), \tag{6.1}$$

where Σ is the experimental covariance matrix.

Given the observed data $\mathbf{d}$, the Bayesian posterior density function of the parameters $\mathbf{c}$ is given by:

$$p(\mathbf{c}|\mathbf{d}) \propto \pi(\mathbf{c}) \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{t}(\mathbf{c}))^T \Sigma^{-1} (\mathbf{d} - \mathbf{t}(\mathbf{c}))\right), \tag{6.2}$$

where $\pi(\mathbf{c})$ is our assumed prior distribution of the parameter $\mathbf{c}$. A $100\alpha\%$ *highest density credible region* R is then a region satisfying:

$$N \int_R \pi(\mathbf{c}) \exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{t}(\mathbf{c}))^T \Sigma^{-1} (\mathbf{d} - \mathbf{t}(\mathbf{c}))\right) = \alpha, \tag{6.3}$$

where N is an appropriate normalisation constant for the posterior. The aim of Bayesian interval estimation is to produce such a region; in order to do so, efficient sampling from the Bayesian posterior is required. Again, this is guaranteed by methods such as Nested Sampling.

6.2 The Monte Carlo replica method in the multivariable case

Let us now describe the application of the Monte Carlo replica method to the multivariable problem presented in Sect. 6.1. Given the observed value of the data $\mathbf{d}$, as usual we

[1]Technically we only require once continuously-differentiable.

construct pseudodata $\mathbf{d}_p$ by sampling from the multivariate normal distribution:

$$\mathbf{d}_p \sim \mathcal{N}(\mathbf{d}, \Sigma). \tag{6.4}$$

We then construct the best-fit theory parameters $\mathbf{c}_p(\mathbf{d}_p)$ by minimising the χ^2-loss to the pseudodata:

$$\mathbf{c}_p(\mathbf{d}_p) = \arg\min_{\mathbf{c}} \, (\mathbf{d}_p - \mathbf{t}(\mathbf{c}))^T \Sigma^{-1} (\mathbf{d}_p - \mathbf{t}(\mathbf{c})). \tag{6.5}$$

This function is possibly multi-valued; for simplicity, we shall assume that it is only *discretely* multi-valued (i.e. we do not have flat directions), and let $\mathbf{c}_p^{(i)}$ for $i = 1, ..., N_{\text{branch}}$ denote the branches of the function. The probability density function of $\mathbf{c}_p(\mathbf{d}_p)$ is then given by:

$$f_{\mathbf{c}_p(\mathbf{d}_p)}(\mathbf{c}) = \int\limits_{\mathbb{R}^{N_{\text{dat}}}} d^{N_{\text{dat}}}\mathbf{d}_p \, \left(\sum_{i=1}^{N_{\text{branch}}(\mathbf{d}_p)} \delta\left(\mathbf{c} - \mathbf{c}_p^{(i)}(\mathbf{d}_p)\right) \right) \exp\left(-\frac{1}{2} (\mathbf{d}_p - \mathbf{d})^T \Sigma^{-1} (\mathbf{d}_p - \mathbf{d}) \right). \tag{6.6}$$

Now, for ease of exposition, we shall assume that $N_{\text{branch}}(\mathbf{d}_p) \equiv 1$; the generalisation to multiple branches is straightforward.

The defining equation Eq. (6.5) implies that $\mathbf{c}_p(\mathbf{d}_p)$ satisfies:

$$\mathbf{0} = \frac{\partial \mathbf{t}^T}{\partial \mathbf{c}}(\mathbf{c}_p(\mathbf{d}_p)) \Sigma^{-1} \left(\mathbf{d}_p - \mathbf{t}(\mathbf{c}_p(\mathbf{d}_p)) \right), \tag{6.7}$$

where the $N_{\text{dat}} \times N_{\text{param}}$ Jacobian matrix $\partial \mathbf{t} / \partial \mathbf{c}$ is given by:

$$\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}} \right)_{ij} = \frac{\partial t_i}{\partial c_j}. \tag{6.8}$$

Suppose that we are given a fixed value of the parameters $\mathbf{c}$, and we wish to determine the posterior distribution function $f_{\mathbf{c}_p(\mathbf{d}_p)}(\mathbf{c})$ at this point. We can centre the pseudodata on the corresponding theory point by making the change of variables:

$$\mathbf{d}_p = \mathbf{t}(\mathbf{c}) + \mathbf{w}, \tag{6.9}$$

which is always valid and has Jacobian matrix given by the identity. The posterior distribution then becomes:

$$f_{\mathbf{c}_p(\mathbf{d}_p)}(\mathbf{c}) = \int\limits_{\mathbb{R}^{N_{\text{dat}}}} d^{N_{\text{dat}}}\mathbf{w} \, \delta\left(\mathbf{c} - \mathbf{c}_p(\mathbf{t}(\mathbf{c}) + \mathbf{w})\right) \exp\left(-\frac{1}{2} (\mathbf{t}(\mathbf{c}) + \mathbf{w} - \mathbf{d})^T \Sigma^{-1} (\mathbf{t}(\mathbf{c}) + \mathbf{w} - \mathbf{d}) \right), \tag{6.10}$$

which simplifies to:

$$\exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)$$
$$\cdot \int_{\mathbb{R}^{N_{\text{dat}}}} d^{N_{\text{dat}}}\mathbf{w}\ \delta\left(\mathbf{c}-\mathbf{c}_p(\mathbf{t}(\mathbf{c})+\mathbf{w})\right)\exp\left(-\frac{1}{2}\mathbf{w}^T\Sigma^{-1}\mathbf{w}+\mathbf{w}^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right) \tag{6.11}$$

The delta function condition is satisfied only if:

$$\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T(\mathbf{c})\Sigma^{-1}\mathbf{w}=0 \qquad \Leftrightarrow \qquad \mathbf{w}\in \text{Ker}\left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}\right). \tag{6.12}$$

However, this is not *sufficient* for the delta function to be satisfied (the condition may lead to a maximum rather than a minimum of the χ^2-statistic); the integration range must be additionally restricted only to values of $\mathbf{w}$ which lead to a minimum at the end of the calculation.

We now proceed to make a change of variables that will allow us to eliminate the delta function. Let $M(\mathbf{c})$ be a matrix whose columns form a basis of the kernel:

$$\text{Ker}\left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}\right), \tag{6.13}$$

and let $M'(\mathbf{c})$ be a matrix whose columns form a basis of the orthogonal space:

$$\text{Ker}\left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}\right)_{\perp}. \tag{6.14}$$

It follows that the matrix $(M|M')$ is invertible, since its columns are linearly independent. We now make the following definition for our change of variables:

$$\mathbf{w}=(M|M')\begin{pmatrix}\mathbf{v}_\perp\\ \mathbf{v}_\parallel\end{pmatrix}=M\mathbf{v}_\perp+M'\mathbf{v}_\parallel. \tag{6.15}$$

Taking the derivative we have $|\det(\partial\mathbf{w}/\partial(\mathbf{v}_\perp,\mathbf{v}_\parallel))|=|\det(M|M')|$. Thus overall we have:

$$|\det(M|M')|\exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)\int d\mathbf{v}_\perp\int d\mathbf{v}_\parallel\ \delta\left(\mathbf{c}-\mathbf{c}_p\left(\mathbf{t}(\mathbf{c})+M\mathbf{v}_\perp+M'\mathbf{v}_\parallel\right)\right)$$
$$\cdot\exp\left(-\frac{1}{2}(M\mathbf{v}_\perp+M'\mathbf{v}_\parallel)^T\Sigma^{-1}(M\mathbf{v}_\perp+M'\mathbf{v}_\parallel)+(M\mathbf{v}_\perp+M'\mathbf{v}_\parallel)^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right). \tag{6.16}$$

The delta function condition is satisfied only if $M'\mathbf{v}_\parallel=\mathbf{0}$, so we can apply this directly to

the exponential, reducing the posterior distribution to:

$$|\det(M|M')| \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)\int d\mathbf{v}_\perp \int d\mathbf{v}_\parallel\ \delta\left(\mathbf{c}-\mathbf{c}_p\left(\mathbf{t}(\mathbf{c})+M\mathbf{v}_\perp+M'\mathbf{v}_\parallel\right)\right)$$

$$\cdot \exp\left(-\frac{1}{2}\mathbf{v}_\perp^T M^T\Sigma^{-1}M\mathbf{v}_\perp+\mathbf{v}_\perp^T M^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right). \tag{6.17}$$

Completing the square in the exponential, the posterior can be rewritten as:

$$|\det(M|M')| \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\left(\Sigma^{-1}-\Sigma^{-1}M(M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}\right)(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)$$

$$\cdot\int d\mathbf{v}_\perp \int d\mathbf{v}_\parallel\ \delta\left(\mathbf{c}-\mathbf{c}_p\left(\mathbf{t}(\mathbf{c})+M\mathbf{v}_\perp+M'\mathbf{v}_\parallel\right)\right)$$

$$\cdot\exp\Bigg(-\frac{1}{2}\left(\mathbf{v}_\perp-(M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)^T$$

$$\cdot M^T\Sigma^{-1}M\left(\mathbf{v}_\perp-(M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)\Bigg). \tag{6.18}$$

We now attempt to remove as much explicit M dependence as possible in this posterior (since of course the final posterior should not depend upon the choice of M). Introducing the standard *Cholesky decomposition* $\Sigma = AA^T$ of the covariance matrix, where A is an invertible matrix, we have that $\Sigma^{-1} = A^{-T}A^{-1}$ and hence:

$$\Sigma^{-1}M(M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1} = A^{-T}\left[(A^{-1}M)((A^{-1}M)^T(A^{-1}M))^{-1}(A^{-1}M)^T\right]A^{-1}. \tag{6.19}$$

We now recall that the matrix $\left[(A^{-1}M)((A^{-1}M)^T(A^{-1}M))^{-1}(A^{-1}M)^T\right]$ is the projector onto the subspace spanned by the columns of $A^{-1}M$. In particular, recalling the definition of M, we have that the columns of:

$$A^T\Sigma^{-1}\frac{\partial\mathbf{t}}{\partial\mathbf{c}} \tag{6.20}$$

span the subspace orthogonal to the subspace spanned by the columns of $A^{-1}M$. It follows that the orthogonal projector $I-(A^{-1}M)((A^{-1}M)^T(A^{-1}M))^{-1}(A^{-1}M)^T$ is given by:

$$\left(A^T\Sigma^{-1}\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\right)\left(\left(A^T\Sigma^{-1}\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\right)^T\left(A^T\Sigma^{-1}\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\right)\right)^{-1}\left(A^T\Sigma^{-1}\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\right)^T. \tag{6.21}$$

Simplifying, this can be reduced to:

$$A^{-1}\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\left(\left(\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\right)^T\Sigma^{-1}\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\right)^{-1}\left(\frac{\partial\mathbf{t}}{\partial\mathbf{c}}\right)^T A^{-T}. \tag{6.22}$$

Hence we have established:

$$\Sigma^{-1} - \Sigma^{-1}M(M^T\Sigma^{-1}M)^{-1}M^T\Sigma^{-1} = \Sigma^{-1}\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^{-1}\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}. \tag{6.23}$$

For ease of notation, we shall define the *tangent inverse covariance matrix* by:

$$\Sigma_t^{-1}(\mathbf{c}) := \Sigma^{-1}\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^{-1}\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}, \tag{6.24}$$

(though note that it is not typically invertible, so we must take care with the notation!) and the *orthogonal inverse covariance matrix* by:

$$\Sigma_o^{-1}(\mathbf{c}) := \Sigma^{-1} - \Sigma^{-1}\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^{-1}\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T\Sigma^{-1}. \tag{6.25}$$

Then we can simplify the posterior distribution to:

$$|\det(M|M')|\exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\Sigma_t^{-1}(\mathbf{c})(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)\int d\mathbf{v}_\perp\int d\mathbf{v}_\parallel\ \delta\left(\mathbf{c}-\mathbf{c}_p\left(\mathbf{t}(\mathbf{c})+M\mathbf{v}_\perp+M'\mathbf{v}_\parallel\right)\right)$$

$$\cdot\exp\left(-\frac{1}{2}\left(\Sigma^{-1}M\mathbf{v}_\perp-\Sigma_o^{-1}(\mathbf{c})(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)^T\Sigma\left(\Sigma^{-1}M\mathbf{v}_\perp-\Sigma_o^{-1}(\mathbf{c})(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)\right). \tag{6.26}$$

Depending on the dimension of the space $\mathrm{Ker}((\partial\mathbf{t}/\partial\mathbf{c})^T\Sigma^{-1})$, there are different sizes of the vector $\mathbf{v}_\parallel$. If $\dim(\mathrm{Ker}((\partial\mathbf{t}/\partial\mathbf{c})^T\Sigma^{-1})) = N_{\mathrm{dat}} - N_{\mathrm{param}}$, which is the typical case, then $\mathbf{v}_\parallel$ has size N_{param} so can absorb all the delta functions. In particular, we should be able to evaluate:

$$\int d\mathbf{v}_\parallel\ \delta\left(\mathbf{c}-\mathbf{c}_p\left(\mathbf{t}(\mathbf{c})+M\mathbf{v}_\perp+M'\mathbf{v}_\parallel\right)\right). \tag{6.27}$$

To do so, consider making the change of variables $\mathbf{c}' = \mathbf{c}_p(\mathbf{t}(\mathbf{c}) + M\mathbf{v}_\perp + M'\mathbf{v}_\parallel)$ in the integral. This implies that $\mathbf{c}'$ is the minimum of the χ^2-statistic on the pseudodata $\mathbf{t}(\mathbf{c}) + M\mathbf{v}_\perp + M'\mathbf{v}_\parallel$, and hence $\mathbf{c}'$ is defined by the implicit equation:

$$\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T(\mathbf{c}')\Sigma^{-1}\left(\mathbf{t}(\mathbf{c})+M\mathbf{v}_\perp+M'\mathbf{v}_\parallel-\mathbf{t}(\mathbf{c}')\right)=\mathbf{0}. \tag{6.28}$$

Recalling the definition of M, we can rewrite this in the form:

$$\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T(\mathbf{c}')\Sigma^{-1}M'\mathbf{v}_\parallel = \left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T(\mathbf{c}')\Sigma^{-1}\left(\mathbf{t}(\mathbf{c})-\mathbf{t}(\mathbf{c}')\right). \tag{6.29}$$

Rearranging, we have:

$$\mathbf{v}_{\|} = \left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}} \right)^T (\mathbf{c}') \Sigma^{-1} M' \right)^{-1} \left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}} \right)^T (\mathbf{c}') \Sigma^{-1} \left(\mathbf{t}(\mathbf{c}) - \mathbf{t}(\mathbf{c}') \right). \tag{6.30}$$

Now writing the identity in the form:

$$I = M(M^T M)^{-1} M^T + M'(M'^T M')^{-1} M'^T, \tag{6.31}$$

using the two projectors on the relevant subspaces, we can rewrite the right hand side as:

$$\mathbf{v}_{\|} = (M'^T M')^{-1} M'^T (\mathbf{t}(\mathbf{c}) - \mathbf{t}(\mathbf{c}')). \tag{6.32}$$

This reveals that the Jacobian of the transformation, evaluated at $\mathbf{c}' = \mathbf{c}$, is:

$$\det \left(\frac{\partial \mathbf{v}_{\|}}{\partial \mathbf{c}'} \right) = \det \left((M'^T M')^{-1} M'^T \frac{\partial \mathbf{t}}{\partial \mathbf{c}} \right). \tag{6.33}$$

Hence the posterior distribution reduces to:

$$|\det(M|M')| \left| \det \left((M'^T M')^{-1} M'^T \frac{\partial \mathbf{t}}{\partial \mathbf{c}} \right) \right| \exp \left(-\frac{1}{2} (\mathbf{d} - \mathbf{t}(\mathbf{c}))^T \Sigma_t^{-1}(\mathbf{c}) (\mathbf{d} - \mathbf{t}(\mathbf{c})) \right) \int d\mathbf{v}_{\perp}$$
$$\cdot \exp \left(-\frac{1}{2} \left(\Sigma^{-1} M \mathbf{v}_{\perp} - \Sigma_o^{-1}(\mathbf{c})(\mathbf{d} - \mathbf{t}(\mathbf{c})) \right)^T \Sigma \left(\Sigma^{-1} M \mathbf{v}_{\perp} - \Sigma_o^{-1}(\mathbf{c})(\mathbf{d} - \mathbf{t}(\mathbf{c})) \right) \right). \tag{6.34}$$

Now without loss of generality, we can choose $(M|M')$ to have orthonormal columns,[2] so that we may further reduce this expression to:

$$\left| \det \left(M'^T \frac{\partial \mathbf{t}}{\partial \mathbf{c}} \right) \right| \exp \left(-\frac{1}{2} (\mathbf{d} - \mathbf{t}(\mathbf{c}))^T \Sigma_t^{-1}(\mathbf{c}) (\mathbf{d} - \mathbf{t}(\mathbf{c})) \right) \int d\mathbf{v}_{\perp}$$
$$\cdot \exp \left(-\frac{1}{2} \left(\Sigma^{-1} M \mathbf{v}_{\perp} - \Sigma_o^{-1}(\mathbf{c})(\mathbf{d} - \mathbf{t}(\mathbf{c})) \right)^T \Sigma \left(\Sigma^{-1} M \mathbf{v}_{\perp} - \Sigma_o^{-1}(\mathbf{c})(\mathbf{d} - \mathbf{t}(\mathbf{c})) \right) \right). \tag{6.35}$$

To simplify things, observe that:

$$\begin{pmatrix} M^T \\ M'^T \end{pmatrix} \begin{pmatrix} M & \frac{\partial \mathbf{t}}{\partial \mathbf{c}} \end{pmatrix} = \begin{pmatrix} M^T M & M^T \frac{\partial \mathbf{t}}{\partial \mathbf{c}} \\ 0 & M'^T \frac{\partial \mathbf{t}}{\partial \mathbf{c}} \end{pmatrix}, \tag{6.36}$$

[2] In fact, the subsequent calculation can be performed without this assumption, but the algebra is significantly trickier.

so that on taking determinants we have:

$$\det\left(M'^T\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right) = \det\begin{pmatrix}M^T\\ M'^T\end{pmatrix}\det\begin{pmatrix}M & \frac{\partial \mathbf{t}}{\partial \mathbf{c}}\end{pmatrix} = \det\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\middle| M\right). \tag{6.37}$$

Hence the Monte Carlo posterior can be rewritten in the form:

$$\left|\det\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\middle| M\right)\right| \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\Sigma_t^{-1}(\mathbf{c})(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)$$
$$\cdot\int d\mathbf{v}_\perp \exp\left(-\frac{1}{2}\left(\Sigma^{-1}M\mathbf{v}_\perp - \Sigma_o^{-1}(\mathbf{c})(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)^T\Sigma\left(\Sigma^{-1}M\mathbf{v}_\perp - \Sigma_o^{-1}(\mathbf{c})(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)\right). \tag{6.38}$$

Some final tidying shows that this scales like the Bayesian posterior, up to a $\mathbf{c}$-dependent factor. Expanding the argument of the exponential inside the orthogonal integral, and removing the $\mathbf{v}_\perp$-independent part, this is equivalent to:

$$\left|\det\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\middle| M\right)\right| \exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\left(\Sigma_t^{-1}+\Sigma_o^{-1}\Sigma\Sigma_o^{-1}\right)(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)$$
$$\cdot\int d\mathbf{v}_\perp \exp\left(-\frac{1}{2}\mathbf{v}_\perp^T M^T\Sigma^{-1}M\mathbf{v}_\perp + \mathbf{v}_\perp^T M^T\Sigma^{-1}\Sigma_o^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right). \tag{6.39}$$

A quick calculation reveals that $\Sigma_t^{-1}+\Sigma_o^{-1}\Sigma\Sigma_o^{-1} = \Sigma^{-1}$, which yields the final form of the Monte Carlo posterior:

$$\left|\det\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\middle| M\right)\right| \int d\mathbf{v}_\perp \exp\left(-\frac{1}{2}\mathbf{v}_\perp^T M^T\Sigma^{-1}M\mathbf{v}_\perp + \mathbf{v}_\perp^T M^T\Sigma^{-1}\Sigma_o^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right)$$
$$\cdot\exp\left(-\frac{1}{2}(\mathbf{d}-\mathbf{t}(\mathbf{c}))^T\Sigma^{-1}(\mathbf{d}-\mathbf{t}(\mathbf{c}))\right) \tag{6.40}$$

It is important to note that the delta function also enforces a restricted range of $\mathbf{v}_\perp$; the integration range here is only over $\mathbf{v}_\perp$ such that the pseudodata $\mathbf{t}(\mathbf{c}) + M\mathbf{v}_\perp$ leads to $\mathbf{c}$ as a minimum of the χ^2-statistic.

We make the key observation that the final distribution is proportional to the Bayesian posterior, except for a $\mathbf{c}$-dependent overall scaling factor. The determinant factor in this scaling factor is easy to compute, if the theory is easily computable (and indeed this must be the case for any Bayesian method anyway). On the other hand, the integral is extremely hard to calculate without numerics, given the complicated integration range.

Special cases. It is also useful to note some special cases where the integral in Eq. (6.40) is analytically tractable. Simple examples include the following:

- **One datapoint, one parameter, quadratic theory.** Consider a single datapoint d with variance σ^2, described by a quadratic theory of one variable, $t(c) = t^{\rm SM} + t^{\rm lin}c + t^{\rm quad}c^2$, as considered in Sect. 4.7. In this case, we have:

$$\frac{\partial t}{\partial c} = t^{\rm lin} + 2ct^{\rm quad} \qquad \Rightarrow \qquad \mathrm{Ker}\left(\frac{1}{\sigma^2}\frac{\partial t}{\partial c}\right) = \begin{cases} 0, & \text{if } c \neq -t^{\rm lin}/2t^{\rm quad}; \\ \mathbb{R}, & \text{if } c = -t^{\rm lin}/2t^{\rm quad}. \end{cases} \tag{6.41}$$

  So provided $c \neq -t^{\rm lin}/2t^{\rm quad}$, we are in the happy case where the dimension of the kernel is zero, and we can apply the above results. The matrix $M(c)$ is empty in this case and the integral over $\mathbf{v}_\perp$ is vacuous; hence the posterior distribution for $c \neq -t^{\rm lin}/2t^{\rm quad}$ reduces simply to:

$$\left|t^{\rm lin} + 2ct^{\rm quad}\right| \exp\left(-\frac{1}{2\sigma^2}(d - t(c))^2\right), \tag{6.42}$$

  which agrees exactly with the result presented in Sect. 4.7.

- **Linear theory.** Consider multiple datapoints $\mathbf{d}$ with covariance Σ, described by a purely linear theory of multiple parameters, $\mathbf{t}(\mathbf{c}) = \mathbf{t}^{\rm SM} + \mathbf{t}^{\rm lin}\mathbf{c}$, with $\mathbf{t}^{\rm lin}$ a $N_{\rm dat} \times N_{\rm param}$ matrix. In this case, we have:

$$\frac{\partial \mathbf{t}}{\partial \mathbf{c}} = \mathbf{t}^{\rm lin} \qquad \Rightarrow \qquad \mathrm{Ker}\left(\left(\frac{\partial \mathbf{t}}{\partial \mathbf{c}}\right)^T \Sigma^{-1}\right) = \{\mathbf{v} : (\mathbf{t}^{\rm lin})^T\Sigma^{-1}\mathbf{v} = \mathbf{0}\}. \tag{6.43}$$

  We note that the relevant kernel here is always independent of $\mathbf{c}$, and hence we may take $M(\mathbf{c})$ independent of $\mathbf{c}$ too. We also note that the tangential covariance matrix is given by:

$$\Sigma_t^{-1} = \Sigma^{-1}\mathbf{t}^{\rm lin}(\mathbf{t}^{\rm linT}\Sigma^{-1}\mathbf{t}^{\rm lin})^{-1}\mathbf{t}^{\rm linT}\Sigma^{-1}, \tag{6.44}$$

  so that $\Sigma_t^{-1}\mathbf{t}^{\rm lin}\mathbf{c} = \Sigma^{-1}\mathbf{t}^{\rm lin}\mathbf{c}$ and $\Sigma_o^{-1}\mathbf{t}^{\rm lin}\mathbf{c} = (\Sigma^{-1} - \Sigma_t^{-1})\mathbf{t}^{\rm lin}\mathbf{c} = \mathbf{0}$. In particular, it follows that the orthogonal integral in Eq. (6.40) is independent of $\mathbf{c}$,[3] and the determinant factor is also independent of $\mathbf{c}$. This leaves just the expected Bayesian posterior:

$$\exp\left(-\frac{1}{2}(\mathbf{d} - \mathbf{t}^{\rm SM} - \mathbf{t}^{\rm lin}\mathbf{c})^T\Sigma^{-1}(\mathbf{d} - \mathbf{t}^{\rm SM} - \mathbf{t}^{\rm lin}\mathbf{c})\right). \tag{6.45}$$

  Hence we have demonstrated that in the purely linear case, the Monte Carlo replica method and the Bayesian method coincide. It should be noted that this is because

[3]Technically, one must additionally confirm that the integration range is independent of $\mathbf{c}$. It is indeed possible to show that the pseudodata $\mathbf{t}^{\rm SM} + \mathbf{t}^{\rm lin}\mathbf{c} + M\mathbf{v}_\perp$ leads to $\mathbf{c}$ as a minimiser of the χ^2 for all $\mathbf{c}$, hence it follows that in fact the integration range is the full space.

of a particularly fortuitous cancellation, however, and that the general behaviour of the Monte Carlo replica method is *not* Bayesian.

- **Pure quadratic theory.** Consider multiple datapoints $\mathbf{d}$ with covariance Σ, described by a purely quadratic theory of one variable, $\mathbf{t}(c) = \mathbf{t}^{\rm SM} + \mathbf{t}^{\rm quad}c^2$, with $\mathbf{t}^{\rm quad} \neq \mathbf{0}$. In this case, we have:

$$\frac{\partial \mathbf{t}}{\partial c} = 2c\mathbf{t}^{\rm quad} \qquad \Rightarrow \qquad \text{Ker}\left(\left(\frac{\partial \mathbf{t}}{\partial c}\right)^T \Sigma^{-1}\right) = \begin{cases} \mathbb{R}^{N_{\rm dat}}, & \text{if } c = 0; \\ \{\mathbf{v} : (\mathbf{t}^{\rm quad})^T \Sigma^{-1} \mathbf{v} = 0\}, & \text{otherwise.} \end{cases} \tag{6.46}$$

So provided $c \neq 0$, we are in the happy case where the dimension of the kernel is $N_{\rm dat} - 1$. Furthermore, we can choose the matrix $M(\mathbf{c})$ independent of $\mathbf{c}$, as in the linear case. Once again, we note that the tangential covariance matrix is c-independent and is given by:

$$\Sigma_t^{-1} = \Sigma^{-1}\mathbf{t}^{\rm quad}(\mathbf{t}^{\rm quad T}\Sigma^{-1}\mathbf{t}^{\rm quad})^{-1}\mathbf{t}^{\rm quad T}\Sigma^{-1}. \tag{6.47}$$

Hence, it follows that $\Sigma_t^{-1}\mathbf{t}^{\rm quad}c^2 = \Sigma^{-1}\mathbf{t}^{\rm quad}c^2$, and so $\Sigma_o^{-1}\mathbf{t}^{\rm quad}c^2 = \mathbf{0}$. In particular, this implies that again the orthogonal integral in Eq. (6.40) is c-independent, and can be simply considered a constant of proportionality.[4] On the other hand, the determinant factor in Eq. (6.40) is *not* c-independent, but is in fact proportional to $|c|$. Overall then, we obtain the posterior:

$$|c| \exp\left(-\frac{1}{2}\left(\mathbf{d} - \mathbf{t}^{\rm SM} - \mathbf{t}^{\rm quad}c^2\right)^T \Sigma^{-1} \left(\mathbf{d} - \mathbf{t}^{\rm SM} - \mathbf{t}^{\rm quad}c^2\right)\right). \tag{6.48}$$

This is equivalent to the Bayesian posterior, up to an overall scale factor.

6.3 Conclusions and future directions

The discussion in the previous section reveals that the Monte Carlo replica method may result in unfaithful error estimates when applied to multivariable problems, particularly when they deviate from the linear case. We already showed in Sect. 4.7 that this can cause serious problems when quadratic corrections dominate in the EFT, but the problem is also present in PDF fitting.

Whilst a considerable amount of the data in modern PDF fits is deep inelastic scattering data, which is linear in the PDFs (we proved factorisation for DIS in the introductory chapter, producing a factorisation theorem of the form $F_2 = \hat{F}_2 \otimes f$ for the structure

[4] Again, technically, one must additionally confirm that the integration range is independent of c. Once again, it is possible to show that $\mathbf{t}^{\rm SM} + \mathbf{t}^{\rm quad}c^2 + M\mathbf{v}_\perp$ leads to c as a minimiser of the χ^2 for all c, and thus for all c the integration range is the full space.

functions), there is a considerable amount of data in which the PDFs enter non-linearly. Both the Drell-Yan data from Chapter 3 and the top production data from Chapter 4 are described by factorisation theorems of the form $\sigma = \hat{\sigma} \otimes f \otimes f$ where the leading contribution is *quadratic* in the PDFs; indeed, the quadratic terms in the PDFs *must* dominate for these datasets because there is no linear contribution. Therefore, it is interesting to ask what effect the use of the Monte Carlo replica method has on global PDF sets which include an increasing amount of hadronic data, involving two protons in the initial state.

Overall, this suggests the need for a new PDF fitting framework beyond the methodology of SIMUnet presented in Chapter 4. In particular, such a framework should be capable of:

- Simultaneously fitting PDFs and theory parameters, in the manner that this thesis has described. This will become increasingly important as we enter the high-luminosity phase of the LHC, as demonstrated in Chapter 3.

- Performing *Bayesian* fits with faithful error estimation for both the PDFs and the theory parameters. The underestimation of errors by the Monte Carlo replica method may lead to false conclusions as increasingly large amounts of data are included in the fits for which the leading contributions are quadratic. This is particularly damaging in the EFT where we may wrongly conclude the existence of New Physics, but it also potentially very damaging in precision SM calculations with PDFs - however, the impact there is yet to be assessed.

We defer the ambitious project of conceiving of, designing and implementing such a global fitting framework to future work.

References

[1] Standard Model Summary Plots February 2022. 2022.

[2] Saranya Samik Ghosh. Highlights from the Compact Muon Solenoid (CMS) Experiment. *Universe*, 5(1):28, 2019.

[3] R. L. Workman et al. Review of Particle Physics. *PTEP*, 2022:083C01, 2022.

[4] Bertrand Delamotte. A hint of renormalization. *Am. J. Phys.*, 72:170–184, 2004.

[5] David J. Gross and Frank Wilczek. Ultraviolet Behavior of Nonabelian Gauge Theories. *Phys. Rev. Lett.*, 30:1343–1346, 1973.

[6] H. David Politzer. Reliable Perturbative Results for Strong Interactions? *Phys. Rev. Lett.*, 30:1346–1349, 1973.

[7] R. P. Feynman. The behavior of hadron collisions at extreme energies. *Conf. Proc. C*, 690905:237–258, 1969.

[8] F. Halzen and Alan D. Martin. *QUARKS AND LEPTONS: AN INTRODUCTORY COURSE IN MODERN PARTICLE PHYSICS*. 1984.

[9] R. Keith Ellis, W. James Stirling, and B. R. Webber. *QCD and collider physics*, volume 8. Cambridge University Press, 2 2011.

[10] Alan D. Martin, W. James Stirling, and R. G. Roberts. Parton distributions of the proton. *Phys. Rev. D*, 50:6734–6752, 1994.

[11] John C. Collins. *Renormalization: An Introduction to Renormalization, The Renormalization Group, and the Operator Product Expansion*, volume 26 of *Cambridge Monographs on Mathematical Physics*. Cambridge University Press, Cambridge, 1986.

[12] Guido Altarelli and G. Parisi. Asymptotic Freedom in Parton Language. *Nucl. Phys. B*, 126:298–318, 1977.

[13] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 3 2014.

[14] John C. Collins, Davison E. Soper, and George F. Sterman. Factorization of Hard Processes in QCD. *Adv. Ser. Direct. High Energy Phys.*, 5:1–91, 1989.

[15] Christian W. Bauer, Dan Pirjol, and Iain W. Stewart. Soft collinear factorization in effective field theory. *Phys. Rev. D*, 65:054022, 2002.

[16] Yuri L. Dokshitzer. Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics. *Sov. Phys. JETP*, 46:641–653, 1977.

[17] V. N. Gribov and L. N. Lipatov. Deep inelastic e p scattering in perturbation theory. *Sov. J. Nucl. Phys.*, 15:438–450, 1972.

[18] A. Vogt, S. Moch, and J. A. M. Vermaseren. The Three-loop splitting functions in QCD: The Singlet case. *Nucl. Phys. B*, 691:129–181, 2004.

[19] S. Moch, J. A. M. Vermaseren, and A. Vogt. The Three loop splitting functions in QCD: The Nonsinglet case. *Nucl. Phys. B*, 688:101–134, 2004.

[20] J. Blümlein, P. Marquard, C. Schneider, and K. Schönwald. The three-loop unpolarized and polarized non-singlet anomalous dimensions from off shell operator matrix elements. *Nucl. Phys. B*, 971:115542, 2021.

[21] S. Moch, B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt. Low moments of the four-loop splitting functions in QCD. *Phys. Lett. B*, 825:136853, 2022.

[22] Daniel de Florian, Germán F. R. Sborlini, and Germán Rodrigo. QED corrections to the Altarelli–Parisi splitting functions. *Eur. Phys. J. C*, 76(5):282, 2016.

[23] Daniel de Florian, Germán F. R. Sborlini, and Germán Rodrigo. Two-loop QED corrections to the Altarelli-Parisi splitting functions. *JHEP*, 10:056, 2016.

[24] Valerio Bertone, Stefano Carrazza, and Juan Rojo. APFEL: A PDF Evolution Library with QED corrections. *Comput. Phys. Commun.*, 185:1647–1668, 2014.

[25] Alessandro Candido, Felix Hekhorn, and Giacomo Magni. EKO: evolution kernel operators. *Eur. Phys. J. C*, 82(10):976, 2022.

[26] Alessandro Candido, Stefano Forte, and Felix Hekhorn. Can $\overline{\text{MS}}$ parton distributions be negative? *JHEP*, 11:129, 2020.

[27] John Collins, Ted C. Rogers, and Nobuo Sato. Positivity and renormalization of parton densities. *Phys. Rev. D*, 105(7):076010, 2022.

[28] Alessandro Candido, Stefano Forte, Tommaso Giani, and Felix Hekhorn. On the positivity of MSbar parton distributions. 7 2023.

[29] H. D. I. Abarbanel, M. L. Goldberger, and S. B. Treiman. Asymptotic properties of electroproduction structure functions. *Phys. Rev. Lett.*, 22:500–502, 1969.

[30] Krzysztof Cichy, Luigi Del Debbio, and Tommaso Giani. Parton distributions from lattice data: the nonsinglet case. *JHEP*, 10:137, 2019.

[31] Richard D. Ball et al. The path to proton structure at 1% accuracy. *Eur. Phys. J. C*, 82(5):428, 2022.

[32] Tie-Jiun Hou et al. New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC. *Phys. Rev. D*, 103(1):014013, 2021.

[33] Stefano Carrazza, Juan M. Cruz-Martinez, and Roy Stegeman. A data-based parametrization of parton distribution functions. *Eur. Phys. J. C*, 82(2):163, 2022.

[34] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[35] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Juan Rojo, and Maria Ubiali. Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties. *JHEP*, 05:075, 2010.

[36] G. D'Agostini. On the use of the covariance matrix to fit correlated data. *Nucl. Instrum. Meth. A*, 346:306–311, 1994.

[37] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015.

[38] Stefano Carrazza and Juan Cruz-Martinez. Towards a new generation of parton densities with deep learning models. *Eur. Phys. J. C*, 79(8):676, 2019.

[39] Luigi Del Debbio, Stefano Forte, Jose I. Latorre, Andrea Piccione, and Joan Rojo. Unbiased determination of the proton structure function F(2)**p with faithful uncertainty estimation. *JHEP*, 03:080, 2005.

[40] Zahari Kassabov, Maeve Madigan, Luca Mantani, James Moore, Manuel Morales Alvarado, Juan Rojo, and Maria Ubiali. The top quark legacy of the LHC Run II for PDF and SMEFT analyses. *JHEP*, 05:205, 2023.

[41] Stefano Forte and Zahari Kassabov. Why α_s cannot be determined from hadronic processes without simultaneously determining the parton distributions. *Eur. Phys. J. C*, 80(3):182, 2020.

[42] Richard D. Ball, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Zahari Kassabov, Juan Rojo, Emma Slade, and Maria Ubiali. Precision determination of the strong coupling constant within a global PDF analysis. *Eur. Phys. J. C*, 78(5):408, 2018.

[43] Tommaso Giani, Giacomo Magni, and Juan Rojo. SMEFiT: a flexible toolbox for global interpretations of particle physics data with effective field theories. *Eur. Phys. J. C*, 83(5):393, 2023.

[44] Matthew McCullough, James Moore, and Maria Ubiali. The dark side of the proton. *JHEP*, 08:019, 2022.

[45] D. H. Rogstad and G. S. Shostak. Gross Properties of Five Scd Galaxies as Determined from 21-CENTIMETER Observations. *Astrophysical Journal*, 176:315, September 1972.

[46] P J E Peebles. Large-scale background temperature and mass fluctuations due to scale-invariant primeval perturbations. *The Astrophysical Journal*, 263, 1982.

[47] Marco Battaglieri et al. US Cosmic Visions: New Ideas in Dark Matter 2017: Community Report. In *U.S. Cosmic Visions: New Ideas in Dark Matter*, 7 2017.

[48] Marco Fabbrichesi, Emidio Gabrielli, and Gaia Lanfranchi. The Dark Photon. 5 2020.

[49] Matt Graham, Christopher Hearty, and Mike Williams. Searches for Dark Photons at Accelerators. *Ann. Rev. Nucl. Part. Sci.*, 71:37–58, 2021.

[50] Graham D. Kribs, David McKeen, and Nirmal Raj. Breaking up the Proton: An Affair with Dark Forces. *Phys. Rev. Lett.*, 126(1):011801, 2021.

[51] A. W. Thomas, X. G. Wang, and A. G. Williams. Constraints on the dark photon from deep inelastic scattering. *Phys. Rev. D*, 105(3):L031901, 2022.

[52] Bin Yan. Probing the dark photon via polarized DIS scattering at the HERA and EIC. 3 2022.

[53] Aneesh Manohar, Paolo Nason, Gavin P. Salam, and Giulia Zanderighi. How bright is the proton? A precise determination of the photon parton distribution function. *Phys. Rev. Lett.*, 117(24):242002, 2016.

[54] Aneesh V. Manohar, Paolo Nason, Gavin P. Salam, and Giulia Zanderighi. The Photon Content of the Proton. *JHEP*, 12:046, 2017.

[55] A. D. Martin, R. G. Roberts, W. J. Stirling, and R. S. Thorne. Parton distributions incorporating qed contributions. *The European Physical Journal C*, 39(2):155–161, Feb 2005.

[56] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, and Juan Rojo. Parton distributions with qed corrections. *Nuclear Physics B*, 877(2):290–320, Dec 2013.

[57] Valerio Bertone, Stefano Carrazza, Nathan P. Hartland, and Juan Rojo. Illuminating the photon content of the proton within a global PDF analysis. *SciPost Phys.*, 5(1):008, 2018.

[58] T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne. QED parton distribution functions in the MSHT20 fit. *Eur. Phys. J. C*, 82(1):90, 2022.

[59] Marco Guzzi, Keping Xie, Tie-Jiun Hou, Pavel Nadolsky, Carl Schmidt, Mengshi Yan, and C. P. Yuan. CTEQ-TEA group updates: Photon PDF and Impact from heavy flavors in the CT18 global analysis. 10 2021.

[60] Bartosz Fornal, Aneesh V. Manohar, and Wouter J. Waalewijn. Electroweak Gauge Boson Parton Distribution Functions. *JHEP*, 05:106, 2018.

[61] Luca Buonocore, Paolo Nason, Francesco Tramontano, and Giulia Zanderighi. Leptons in the proton. *JHEP*, 08(08):019, 2020.

[62] Luca Buonocore, Ulrich Haisch, Paolo Nason, Francesco Tramontano, and Giulia Zanderighi. Lepton-Quark Collisions at the Large Hadron Collider. *Phys. Rev. Lett.*, 125(23):231804, 2020.

[63] Admir Greljo and Nudzeim Selimovic. Lepton-Quark Fusion at Hadron Colliders, precisely. *JHEP*, 03:279, 2021.

[64] Luca Buonocore, Paolo Nason, Francesco Tramontano, and Giulia Zanderighi. Photon and leptons induced processes at the LHC. *JHEP*, 12:073, 2021.

[65] L. A. Harland-Lang. Physics with leptons and photons at the LHC. *Phys. Rev. D*, 104(7):073002, 2021.

[66] Edmond L. Berger, Pavel M. Nadolsky, Fredrick I. Olness, and Jon Pumplin. Light gluino constituents of hadrons and a global analysis of hadron scattering data. *Phys. Rev. D*, 71:014007, 2005.

[67] Edmond L. Berger, Marco Guzzi, Hung-Liang Lai, Pavel M. Nadolsky, and Fredrick I. Olness. Constraints on color-octet fermions from a global parton distribution analysis. *Phys. Rev. D*, 82:114023, 2010.

[68] Diego Becciolini, Marc Gillioz, Marco Nardecchia, Francesco Sannino, and Michael Spannowsky. Constraining new colored matter from the ratio of 3 to 2 jets cross sections at the LHC. *Phys. Rev. D*, 91(1):015010, 2015. [Addendum: Phys.Rev.D 92, 079905 (2015)].

[69] Valerio Bertone, Stefano Carrazza, Davide Pagani, and Marco Zaro. On the Impact of Lepton PDFs. *JHEP*, 11:194, 2015.

[70] Fabio Maltoni, Giovanni Ridolfi, and Maria Ubiali. b-initiated processes at the LHC: a reappraisal. *JHEP*, 07:022, 2012. [Erratum: JHEP 04, 095 (2013)].

[71] Valerio Bertone, Alexandre Glazov, Alexander Mitov, Andrew Papanastasiou, and Maria Ubiali. Heavy-flavor parton distributions without heavy-flavor matching prescriptions. *JHEP*, 04:046, 2018.

[72] A. De Rujula, R. Petronzio, and A. Savoy-Navarro. Radiative Corrections to High-Energy Neutrino Scattering. *Nucl. Phys. B*, 154:394–426, 1979.

[73] J. Kripfganz and H. Perlt. Electroweak Radiative Corrections and Quark Mass Singularities. *Z. Phys. C*, 41:319–321, 1988.

[74] J. Blumlein. Leading Log Radiative Corrections to Deep Inelastic Neutral and Charged Current Scattering at HERA. *Z. Phys. C*, 47:89–94, 1990.

[75] Andy Buckley, James Ferrando, Stephen Lloyd, Karl Nordström, Ben Page, Martin Rüfenacht, Marek Schönherr, and Graeme Watt. LHAPDF6: parton density access in the LHC precision era. *Eur. Phys. J. C*, 75:132, 2015.

[76] Philip Ilten, Yotam Soreq, Mike Williams, and Wei Xue. Serendipity in dark photon searches. *JHEP*, 06:004, 2018.

[77] Rabah Abdul Khalek, Shaun Bailey, Jun Gao, Lucian Harland-Lang, and Juan Rojo. Towards Ultimate Parton Distributions at the High-Luminosity LHC. *Eur. Phys. J. C*, 78(11):962, 2018.

[78] Bogdan A. Dobrescu and Claudia Frugiuele. Hidden GeV-scale interactions of quarks. *Phys. Rev. Lett.*, 113:061801, 2014.

[79] Jeff A. Dror, Robert Lasenby, and Maxim Pospelov. New constraints on light vectors coupled to anomalous currents. *Phys. Rev. Lett.*, 119(14):141803, 2017.

[80] Ahmed Ismail and Andrey Katz. Anomalous Z-prime and diboson resonances at the LHC. *JHEP*, 04:122, 2018.

[81] Jeff A. Dror, Robert Lasenby, and Maxim Pospelov. Light vectors coupled to bosonic currents. *Phys. Rev. D*, 99(5):055016, 2019.

[82] Bogdan A. Dobrescu and Felix Yu. Dijet and electroweak limits on a Z' boson coupled to quarks. 12 2021.

[83] Michael Duerr, Pavel Fileviez Perez, and Mark B. Wise. Gauge Theory for Baryon and Lepton Numbers with Leptoquarks. *Phys. Rev. Lett.*, 110:231801, 2013.

[84] Pavel Fileviez Perez, Sebastian Ohmer, and Hiren H. Patel. Minimal Theory for Lepto-Baryons. *Phys. Lett. B*, 735:283–287, 2014.

[85] Pavel Fileviez Pérez, Elliot Golias, Rui-Hao Li, Clara Murgui, and Alexis D. Plascencia. Anomaly-free dark matter models. *Phys. Rev. D*, 100(1):015017, 2019.

[86] Pavel Fileviez Pérez and Alexis D. Plascencia. Electric dipole moments, new forces and dark matter. *JHEP*, 03:185, 2021.

[87] Pavel Fileviez Perez and Alexis D. Plascencia. Theory of Dirac Dark Matter: Higgs Decays and EDMs. 12 2021.

[88] Pavel Fileviez Pérez, Elliot Golias, Clara Murgui, and Alexis D. Plascencia. The Higgs and leptophobic force at the LHC. *JHEP*, 07:087, 2020.

[89] John Preskill. Gauge anomalies in an effective field theory. *Annals Phys.*, 210:323–379, 1991.

[90] M. Acciarri et al. Search for new particles in hadronic events with isolated photons. *Phys. Lett. B*, 388:409–418, 1996.

[91] B. Adeva et al. Search for narrow high mass resonances in radiative decays of the Z0. *Phys. Lett. B*, 262:155–162, 1991.

[92] O. Adriani et al. Isolated hard photon emission in hadronic Z0 decays. *Phys. Lett. B*, 292:472–484, 1992.

[93] Christopher D. Carone and Hitoshi Murayama. Possible light U(1) gauge boson coupled to baryon number. *Phys. Rev. Lett.*, 74:3122–3125, 1995.

[94] Alfredo Aranda and Christopher D. Carone. Limits on a light leptophobic gauge boson. *Phys. Lett. B*, 443:352–358, 1998.

[95] H. Albrecht et al. An Upper Limit for Two Jet Production in Direct Υ (1s) Decays. *Z. Phys. C*, 31:181, 1986.

[96] Bogdan A. Dobrescu and Felix Yu. Coupling-Mass Mapping of Dijet Peak Searches. *Phys. Rev. D*, 88(3):035021, 2013. [Erratum: Phys.Rev.D 90, 079901 (2014)].

[97] Chase Shimmin and Daniel Whiteson. Boosting low-mass hadronic resonances. *Phys. Rev. D*, 94(5):055001, 2016.

[98] Morad Aaboud et al. Search for low-mass resonances decaying into two jets and produced in association with a photon using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett. B*, 795:56–75, 2019.

[99] Albert M Sirunyan et al. Search for Low-Mass Quark-Antiquark Resonances Produced in Association with a Photon at $\sqrt{s}$ =13 TeV. *Phys. Rev. Lett.*, 123(23):231803, 2019.

[100] Albert M Sirunyan et al. Search for low mass vector resonances decaying into quark-antiquark pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D*, 100(11):112007, 2019.

[101] Stefano Carrazza, Celine Degrande, Shayan Iranipour, Juan Rojo, and Maria Ubiali. Can New Physics hide inside the proton? *Phys. Rev. Lett.*, 123(13):132001, 2019.

[102] Admir Greljo, Shayan Iranipour, Zahari Kassabov, Maeve Madigan, James Moore, Juan Rojo, Maria Ubiali, and Cameron Voisey. Parton distributions in the SMEFT from high-energy Drell-Yan tails. *JHEP*, 07:122, 2021.

[103] Shayan Iranipour and Maria Ubiali. A new generation of simultaneous fits to LHC data using deep learning. *JHEP*, 05:032, 2022.

[104] Albert M Sirunyan et al. Measurement of the differential Drell-Yan cross section in proton-proton collisions at $\sqrt{\mathrm{s}} = 13$ TeV. *JHEP*, 12:059, 2019.

[105] Claude Duhr, Falko Dulat, and Bernhard Mistlberger. Drell-Yan Cross Section to Third Order in the Strong Coupling Constant. *Phys. Rev. Lett.*, 125(17):172001, 2020.

[106] Claude Duhr and Bernhard Mistlberger. Lepton-pair production at hadron colliders at $\mathrm{N^3LO}$ in QCD. 11 2021.

[107] Nathan P. Hartland, Fabio Maltoni, Emanuele R. Nocera, Juan Rojo, Emma Slade, Eleni Vryonidou, and Cen Zhang. A Monte Carlo global analysis of the Standard Model Effective Field Theory: the top quark sector. *JHEP*, 04:100, 2019.

[108] Fabian Esser, Maeve Madigan, Veronica Sanz, and Maria Ubiali. On the coupling of axion-like particles to the top quark. 3 2023.

[109] Maeve Madigan and James Moore. Parton Distributions in the SMEFT from high-energy Drell-Yan tails. *PoS*, EPS-HEP2021:424, 2022.

[110] Alexey A. Petrov and Andrew E. Blechman. *Effective Field Theories*. WSP, 2016.

[111] Brian Henning, Xiaochuan Lu, and Hitoshi Murayama. How to use the Standard Model effective field theory. *JHEP*, 01:023, 2016.

[112] Aneesh V. Manohar. Introduction to Effective Field Theories. 4 2018.

[113] Ilaria Brivio and Michael Trott. The Standard Model as an Effective Field Theory. *Phys. Rept.*, 793:1–98, 2019.

[114] B. Grzadkowski, M. Iskrzynski, M. Misiak, and J. Rosiek. Dimension-Six Terms in the Standard Model Lagrangian. *JHEP*, 10:085, 2010.

[115] Jun Gao, Lucian Harland-Lang, and Juan Rojo. The Structure of the Proton in the LHC Precision Era. *Phys. Rept.*, 742:1–121, 2018.

[116] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015.

[117] Richard D. Ball et al. Parton distributions from high-precision collider data. *Eur. Phys. J.*, C77(10):663, 2017.

[118] S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne. Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs. 12 2020.

[119] Rabah Abdul Khalek, Shaun Bailey, Jun Gao, Lucian Harland-Lang, and Juan Rojo. Towards Ultimate Parton Distributions at the High-Luminosity LHC. *Eur. Phys. J. C*, 78(11):962, 2018.

[120] Marco Farina, Giuliano Panico, Duccio Pappadopulo, Joshua T. Ruderman, Riccardo Torre, and Andrea Wulzer. Energy helps accuracy: electroweak precision tests at hadron colliders. *Phys. Lett.*, B772:210–215, 2017.

[121] Michael E. Peskin and Tatsu Takeuchi. Estimation of oblique electroweak corrections. *Phys. Rev.*, D46:381–409, 1992.

[122] Guido Altarelli, Riccardo Barbieri, and S. Jadach. Toward a model independent analysis of electroweak data. *Nucl. Phys.*, B369:3–32, 1992. [Erratum: Nucl. Phys.B376,444(1992)].

[123] Riccardo Barbieri, Alex Pomarol, Riccardo Rattazzi, and Alessandro Strumia. Electroweak symmetry breaking after LEP-1 and LEP-2. *Nucl. Phys.*, B703:127–146, 2004.

[124] Christoph Englert, Gian F. Giudice, Admir Greljo, and Matthew Mccullough. The $\hat{H}$-Parameter: An Oblique Higgs View. *JHEP*, 09:041, 2019.

[125] Ilaria Brivio, Yun Jiang, and Michael Trott. The SMEFTsim package, theory and tools. *JHEP*, 12:070, 2017.

[126] Admir Greljo and David Marzocca. High-p_T dilepton tails and flavor physics. *Eur. Phys. J.*, C77(8):548, 2017.

[127] Riccardo Torre, Lorenzo Ricci, and Andrea Wulzer. On the W&Y interpretation of high-energy Drell-Yan measurements. *JHEP*, 02:144, 2021.

[128] Georges Aad et al. Measurement of the double-differential high-mass Drell-Yan cross section in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 08:009, 2016.

[129] Vardan Khachatryan et al. Measurements of differential and double-differential Drell-Yan cross sections in proton-proton collisions at 8 TeV. *Eur. Phys. J.*, C75(4):147, 2015.

[130] Ferran Faura, Shayan Iranipour, Emanuele R. Nocera, Juan Rojo, and Maria Ubiali. The Strangest Proton? *Eur. Phys. J. C*, 80(12):1168, 2020.

[131] H. Abramowicz et al. Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data. *Eur. Phys. J.*, C75(12):580, 2015.

[132] R. S. Towell et al. Improved measurement of the anti-d / anti-u asymmetry in the nucleon sea. *Phys. Rev.*, D64:052002, 2001.

[133] J. C. Webb et al. Absolute Drell-Yan dimuon cross-sections in 800 GeV / c pp and pd collisions. 2003.

[134] Jason C. Webb. *Measurement of continuum dimuon production in 800-GeV/C proton nucleon collisions*. PhD thesis, New Mexico State U., 2003.

[135] G. Moreno et al. Dimuon production in proton - copper collisions at $\sqrt{s} = 38.8$-GeV. *Phys. Rev. D*, 43:2815–2836, 1991.

[136] Timo Antero Aaltonen et al. Measurement of $d\sigma/dy$ of Drell-Yan e^+e^- pairs in the Z Mass Region from $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV. *Phys. Lett.*, B692:232–239, 2010.

[137] V. M. Abazov et al. Measurement of the Shape of the Boson Rapidity Distribution for $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ Events Produced at $\sqrt{s}$ of 1.96-TeV. *Phys. Rev.*, D76:012003, 2007.

[138] Victor Mukhamedovich Abazov et al. Measurement of the Muon Charge Asymmetry in $p\bar{p} \to$ W+X $\to \mu\nu$ + X Events at $\sqrt{s}$=1.96 TeV. *Phys. Rev. D*, 88:091102, 2013.

[139] Georges Aad et al. Measurement of the inclusive $W^{\pm}$ and Z/gamma cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Phys. Rev.*, D85:072004, 2012.

[140] Georges Aad et al. Measurement of the low-mass Drell-Yan differential cross section at $\sqrt{s} = 7$ TeV using the ATLAS detector. *JHEP*, 06:112, 2014.

[141] Morad Aaboud et al. Precision measurement and interpretation of inclusive W^+ , W^- and Z/γ^* production cross sections with the ATLAS detector. *Eur. Phys. J.*, C77(6):367, 2017.

[142] Georges Aad et al. Measurement of the production of a W boson in association with a charm quark in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *JHEP*, 05:068, 2014.

[143] Georges Aad et al. Measurement of the transverse momentum and ϕ^*_η distributions of Drell–Yan lepton pairs in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Eur. Phys. J.*, C76(5):291, 2016.

[144] Morad Aaboud et al. Measurement of differential cross sections and W^+/W^- cross-section ratios for W boson production in association with jets at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 05:077, 2018. [Erratum: JHEP 10, 048 (2020)].

[145] Serguei Chatrchyan et al. Measurement of the Electron Charge Asymmetry in Inclusive W Production in pp Collisions at $\sqrt{s} = 7$ TeV. *Phys. Rev. Lett.*, 109:111806, 2012.

[146] Serguei Chatrchyan et al. Measurement of Associated W + Charm Production in pp Collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 02:013, 2014.

[147] Vardan Khachatryan et al. Measurement of the Z boson differential cross section in transverse momentum and rapidity in proton–proton collisions at 8 TeV. *Phys. Lett.*, B749:187–209, 2015.

[148] Vardan Khachatryan et al. Measurement of the differential cross section and charge asymmetry for inclusive pp $\rightarrow \mathrm{W}^{\pm} + X$ production at $\sqrt{s} = 8$ TeV. *Eur. Phys. J. C*, 76(8):469, 2016.

[149] Albert M Sirunyan et al. Measurement of associated production of a W boson and a charm quark in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Eur. Phys. J. C*, 79(3):269, 2019.

[150] R Aaij et al. Inclusive W and Z production in the forward region at $\sqrt{s} = 7$ TeV. *JHEP*, 06:058, 2012.

[151] Roel Aaij et al. Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 08:039, 2015.

[152] R Aaij et al. Measurement of the cross-section for $Z \rightarrow e^+e^-$ production in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 02:106, 2013.

[153] Roel Aaij et al. Measurement of forward W and Z boson production in pp collisions at $\sqrt{s} = 8$ TeV. *JHEP*, 01:155, 2016.

[154] Victor Mukhamedovich Abazov et al. Measurement of the electron charge asymmetry in $\boldsymbol{p\bar{p} \rightarrow W + X \rightarrow e\nu + X}$ decays in $\boldsymbol{p\bar{p}}$ collisions at $\boldsymbol{\sqrt{s} = 1.96}$ TeV. *Phys. Rev. D*, 91(3):032007, 2015. [Erratum: Phys.Rev.D 91, 079901 (2015)].

[155] Georges Aad et al. Measurement of the high-mass Drell–Yan differential cross-section in pp collisions at sqrt(s)=7 TeV with the ATLAS detector. *Phys. Lett.*, B725:223–242, 2013.

[156] Serguei Chatrchyan et al. Measurement of the Differential and Double-Differential Drell-Yan Cross Sections in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 12:030, 2013.

[157] Albert M Sirunyan et al. Measurement of the differential Drell-Yan cross section in proton-proton collisions at $\sqrt{\mathrm{s}} = 13$ TeV. *JHEP*, 12:059, 2019.

[158] Stefano Forte, Eric Laenen, Paolo Nason, and Juan Rojo. Heavy quarks in deep-inelastic scattering. *Nuclear Physics B*, 834(1-2):116–162, Jul 2010.

[159] Valerio Bertone, Stefano Carrazza, and Juan Rojo. Apfel: A pdf evolution library with qed corrections. *Computer Physics Communications*, 185(6):1647–1668, Jun 2014.

[160] Valerio Bertone, Stefano Carrazza, and Nathan P. Hartland. APFELgrid: a high performance tool for parton density determinations. *Comput. Phys. Commun.*, 212:205–209, 2017.

[161] John Campbell and Tobias Neumann. Precision Phenomenology with MCFM. *JHEP*, 12:034, 2019.

[162] R. Frederix, S. Frixione, V. Hirschi, D. Pagani, H. S. Shao, and M. Zaro. The automation of next-to-leading order electroweak calculations. *JHEP*, 07:185, 2018.

[163] Tancredi Carli, Dan Clements, Amanda Cooper-Sarkar, Claire Gwenlan, Gavin P. Salam, Frank Siegert, Pavel Starovoitov, and Mark Sutton. A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project. *Eur. Phys. J. C*, 66:503–524, 2010.

[164] Massimiliano Grazzini, Stefan Kallweit, and Marius Wiesemann. Fully differential nnlo computations with matrix. *The European Physical Journal C*, 78(7), Jun 2018.

[165] Ye Li and Frank Petriello. Combining QCD and electroweak corrections to dilepton production in FEWZ. *Phys. Rev. D*, 86:094034, 2012.

[166] Claude Duhr, Falko Dulat, and Bernhard Mistlberger. Charged current Drell-Yan production at N^3LO. *JHEP*, 11:143, 2020.

[167] Richard D. Ball et al. Parton Distribution Benchmarking with LHC Data. *JHEP*, 04:125, 2013.

[168] Simone Lionetti, Richard D. Ball, Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Juan Rojo, and Maria Ubiali. Precision determination of α_s using an unbiased global NLO parton set. *Phys. Lett. B*, 701:346–352, 2011.

[169] Richard D. Ball, Valerio Bertone, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Simone Lionetti, Juan Rojo, and Maria Ubiali. Precision NNLO determination of $\alpha_s(M_Z)$ using an unbiased global parton set. *Phys. Lett. B*, 707:66–71, 2012.

[170] Georges Aad et al. Search for a heavy charged boson in events with a charged lepton and missing transverse momentum from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D*, 100(5):052013, 2019.

[171] Gino Isidori, Giovanni Ridolfi, and Alessandro Strumia. On the metastability of the standard model vacuum. *Nucl. Phys. B*, 609:387–409, 2001.

[172] Dario Buttazzo, Giuseppe Degrassi, Pier Paolo Giardino, Gian F. Giudice, Filippo Sala, Alberto Salvio, and Alessandro Strumia. Investigating the near-criticality of the Higgs boson. *JHEP*, 12:089, 2013.

[173] Luca Di Luzio, Gino Isidori, and Giovanni Ridolfi. Stability of the electroweak ground state in the Standard Model and its extensions. *Phys. Lett. B*, 753:150–160, 2016.

[174] M. Arneodo et al. Accurate measurement of F_2^d/F_2^p and $R_d - R_p$. *Nucl. Phys.*, B487:3–26, 1997.

[175] M. Arneodo et al. Measurement of the proton and deuteron structure functions, F_2^p and F_2^d, and of the ratio σ_L/σ_T. *Nucl. Phys.*, B483:3–43, 1997.

[176] L. W. Whitlow, E. M. Riordan, S. Dasu, Stephen Rock, and A. Bodek. Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross-sections. *Phys. Lett.*, B282:475–482, 1992.

[177] A. C. Benvenuti et al. A High Statistics Measurement of the Proton Structure Functions $F_2(x, Q^2)$ and R from Deep Inelastic Muon Scattering at High Q^2. *Phys. Lett.*, B223:485, 1989.

[178] G. Onengut et al. Measurement of nucleon structure functions in neutrino scattering. *Phys. Lett.*, B632:65–75, 2006.

[179] M. Goncharov et al. Precise measurement of dimuon production cross-sections in ν_μFe and $\bar{\nu}_\mu$Fe deep inelastic scattering at the Tevatron. *Phys. Rev.*, D64:112006, 2001.

[180] David Alexander Mason. Measurement of the strange - antistrange asymmetry at NLO in QCD from NuTeV dimuon data. FERMILAB-THESIS-2006-01.

[181] H. Abramowicz et al. Combination and QCD analysis of charm and beauty production cross-section measurements in deep inelastic *ep* scattering at HERA. *Eur. Phys. J.*, C78(6):473, 2018.

[182] A. Abulencia et al. Measurement of the Inclusive Jet Cross Section using the k_{T} algorithm in $p\overline{p}$ Collisions at $\sqrt{s}$=1.96 TeV with the CDF II Detector. *Phys. Rev.*, D75:092006, 2007.

[183] Serguei Chatrchyan et al. Measurement of the muon charge asymmetry in inclusive pp to WX production at $\sqrt{s} = 7$ TeV and an improved determination of light parton distribution functions. *Phys.Rev.*, D90:032004, 2014.

[184] Roel Aaij et al. Measurement of forward $\mathrm{Z} \to \mathrm{e}^+\mathrm{e}^-$ production at $\sqrt{s} = 8$ TeV. *JHEP*, 05:109, 2015.

[185] M. Aaboud et al. Measurement of the Drell-Yan triple-differential cross section in pp collisions at $\sqrt{s} = 8$ TeV. *JHEP*, 12:059, 2017.

[186] Georges Aad et al. Measurement of the cross-section and charge asymmetry of W bosons produced in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 79(9):760, 2019.

[187] Roel Aaij et al. Measurement of forward $W \to e\nu$ production in pp collisions at $\sqrt{s} = 8\,\text{TeV}$. *JHEP*, 10:030, 2016.

[188] Georges Aad et al. Measurement of $W^{\pm}$ and Z-boson production cross sections in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett.*, B759:601–621, 2016.

[189] Roel Aaij et al. Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 09:136, 2016.

[190] Albert M Sirunyan et al. Measurement of the differential cross sections for the associated production of a W boson and jets in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D*, 96(7):072005, 2017.

[191] Georges Aad et al. Measurement of inclusive jet and dijet production in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector. *Phys. Rev. D*, 86:014022, 2012.

[192] Georges Aad et al. Measurement of the inclusive jet cross section in pp collisions at $\sqrt{s}$=2.76 TeV and comparison to the inclusive jet cross section at $\sqrt{s}$=7 TeV using the ATLAS detector. *Eur.Phys.J.*, C73:2509, 2013.

[193] Georges Aad et al. Measurement of the inclusive jet cross-section in proton-proton collisions at $\sqrt{s} = 7$ TeV using 4.5 fb^{-1} of data with the ATLAS detector. *JHEP*, 02:153, 2015. [Erratum: JHEP09,141(2015)].

[194] Serguei Chatrchyan et al. Measurements of Differential Jet Cross Sections in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV with the CMS Detector. *Phys. Rev. D*, 87(11):112002, 2013. [Erratum: Phys.Rev.D 87, 119902 (2013)].

[195] Vardan Khachatryan et al. Measurement of the inclusive jet cross section in pp collisions at $\sqrt{s} = 2.76\,\text{TeV}$. *Eur. Phys. J.*, C76(5):265, 2016.

[196] Morad Aaboud et al. Measurement of the inclusive jet cross-sections in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 09:020, 2017.

[197] Vardan Khachatryan et al. Measurement and QCD analysis of double-differential inclusive jet cross sections in pp collisions at $\sqrt{s} = 8$ TeV and cross section ratios to 2.76 and 7 TeV. *JHEP*, 03:156, 2017.

[198] Georges Aad et al. Measurement of the inclusive isolated prompt photon cross section in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 08:005, 2016.

[199] Morad Aaboud et al. Measurement of the cross section for inclusive isolated-photon production in *pp* collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector. *Phys. Lett. B*, 770:473–493, 2017.

[200] John Ellis, Maeve Madigan, Ken Mimasu, Veronica Sanz, and Tevong You. Top, Higgs, Diboson and Electroweak Fit to the Standard Model Effective Field Theory. 12 2020.

[201] Jacob J. Ethier, Giacomo Magni, Fabio Maltoni, Luca Mantani, Emanuele R. Nocera, Juan Rojo, Emma Slade, Eleni Vryonidou, and Cen Zhang. Combined SMEFT interpretation of Higgs, diboson, and top quark data from the LHC. *JHEP*, 11:089, 2021.

[202] Georges Aad et al. Measurements of top-quark pair differential cross-sections in the lepton+jets channel in *pp* collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector. *Eur. Phys. J. C*, 76(10):538, 2016.

[203] Michał Czakon, Nathan P. Hartland, Alexander Mitov, Emanuele R. Nocera, and Juan Rojo. Pinning down the large-x gluon with NNLO top-quark pair differential distributions. *JHEP*, 04:044, 2017.

[204] Georges Aad et al. Measurement of the $t\bar{t}$ production cross-section using $e\mu$ events with b-tagged jets in pp collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS detector. *Eur. Phys. J. C*, 74(10):3109, 2014. [Addendum: Eur.Phys.J.C 76, 642 (2016)].

[205] Morad Aaboud et al. Measurement of top quark pair differential cross-sections in the dilepton channel in *pp* collisions at $\sqrt{s} = 7$ and 8 TeV with ATLAS. *Phys. Rev. D*, 94(9):092003, 2016. [Addendum: Phys.Rev.D 101, 119901 (2020)].

[206] Morad Aaboud et al. Measurement of the inclusive and fiducial $t\bar{t}$ production cross-sections in the lepton+jets channel in *pp* collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 78:487, 2018.

[207] Georges Aad et al. Measurement of the $t\bar{t}$ production cross-section and lepton differential distributions in $e\mu$ dilepton events from *pp* collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector. *Eur. Phys. J. C*, 80(6):528, 2020.

[208] Georges Aad et al. Measurements of top-quark pair single- and double-differential cross-sections in the all-hadronic channel in *pp* collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector. *JHEP*, 01:033, 2021.

[209] Georges Aad et al. Measurement of the $t\bar{t}$ production cross-section in the lepton+jets channel at $\sqrt{s} = 13$ TeV with the ATLAS experiment. *Phys. Lett. B*, 810:135797, 2020.

[210] Georges Aad et al. Measurements of top-quark pair differential and double-differential cross-sections in the ℓ+jets channel with pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector. *Eur. Phys. J. C*, 79(12):1028, 2019. [Erratum: Eur.Phys.J.C 80, 1092 (2020)].

[211] A. M. Sirunyan et al. Measurement of the inclusive $\mathrm{t\bar{t}}$ cross section in pp collisions at $\sqrt{s} = 5.02$ TeV using final states with at least one charged lepton. *JHEP*, 03:115, 2018.

[212] Simon Spannagel. Top quark mass measurements with the CMS experiment at the LHC. *PoS*, DIS2016:150, 2016.

[213] Albert M Sirunyan et al. Measurement of double-differential cross sections for top quark pair production in pp collisions at $\sqrt{s} = 8$ TeV and impact on parton distribution functions. *Eur. Phys. J. C*, 77(7):459, 2017.

[214] Vardan Khachatryan et al. Measurement of the differential cross section for top quark pair production in pp collisions at $\sqrt{s} = 8\,\mathrm{TeV}$. *Eur. Phys. J. C*, 75(11):542, 2015.

[215] Vardan Khachatryan et al. Measurement of the top quark pair production cross section in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 116(5):052002, 2016.

[216] Albert M Sirunyan et al. Measurements of $\mathrm{t\bar{t}}$ differential cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV using events containing two leptons. *JHEP*, 02:149, 2019.

[217] Armen Tumasyan et al. Measurement of differential $\mathrm{t\bar{t}}$ production cross sections in the full kinematic range using lepton+jets events from proton-proton collisions at $\sqrt{s} = 13$ TeV. 8 2021.

[218] Georges Aad et al. Measurements of the charge asymmetry in top-quark pair production in the dilepton final state at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Phys. Rev. D*, 94(3):032006, 2016.

[219] Evidence for the charge asymmetry in $pp \to t\bar{t}$ production at $\sqrt{s} = 13$ TeV with the ATLAS detector. 8 2022.

[220] Vardan Khachatryan et al. Measurements of $t\bar{t}$ charge asymmetry using dilepton final states in pp collisions at $\sqrt{s} = 8$ TeV. *Phys. Lett. B*, 760:365–386, 2016.

[221] Measurement of the ttbar charge asymmetry in highly boosted events in the single-lepton channel at 13 TeV. Technical report, CERN, Geneva, 2022.

[222] Morad Aaboud et al. Combination of inclusive and differential $\mathrm{t\bar{t}}$ charge asymmetry measurements using ATLAS and CMS data at $\sqrt{s} = 7$ and 8 TeV. *JHEP*, 04:033, 2018.

[223] Georges Aad et al. Combination of the W boson polarization measurements in top quark decays using ATLAS and CMS data at $\sqrt{s} = 8$ TeV. *JHEP*, 08(08):051, 2020.

[224] Measurement of the polarisation of W bosons produced in top-quark decays using di-lepton events at $\sqrt{s} = 13$ TeV with the ATLAS experiment. 2022.

[225] Georges Aad et al. Measurement of the $t\bar{t}W$ and $t\bar{t}Z$ production cross sections in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 11:172, 2015.

[226] Morad Aaboud et al. Measurement of the $t\bar{t}Z$ and $t\bar{t}W$ cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D*, 99(7):072009, 2019.

[227] Georges Aad et al. Measurements of the inclusive and differential production cross sections of a top-quark–antiquark pair in association with a Z boson at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 81(8):737, 2021.

[228] Vardan Khachatryan et al. Observation of top quark pairs produced in association with a vector boson in pp collisions at $\sqrt{s} = 8$ TeV. *JHEP*, 01:096, 2016.

[229] Albert M Sirunyan et al. Measurement of the cross section for top quark pair production in association with a W or Z boson in proton-proton collisions at $\sqrt{s} =$ 13 TeV. *JHEP*, 08:011, 2018.

[230] Albert M Sirunyan et al. Measurement of top quark pair production in association with a Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 03:056, 2020.

[231] Morad Aaboud et al. Measurement of the $t\bar{t}\gamma$ production cross section in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 11:086, 2017.

[232] Albert M Sirunyan et al. Measurement of the semileptonic $\mathrm{t\bar{t}} + \gamma$ production cross section in pp collisions at $\sqrt{s} = 8$ TeV. *JHEP*, 10:006, 2017.

[233] Georges Aad et al. Evidence for $t\bar{t}t\bar{t}$ production in the multilepton final state in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 80(11):1085, 2020.

[234] Georges Aad et al. Measurement of the $t\bar{t}t\bar{t}$ production cross section in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP*, 11:118, 2021.

[235] Morad Aaboud et al. Measurements of inclusive and differential fiducial cross-sections of $t\bar{t}$ production with additional heavy-flavour jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector. *JHEP*, 04:046, 2019.

[236] Albert M Sirunyan et al. Search for production of four top quarks in final states with same-sign or multiple leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Eur. Phys. J. C*, 80(2):75, 2020.

[237] Albert M Sirunyan et al. Search for the production of four top quarks in the single-lepton and opposite-sign dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 11:082, 2019.

[238] Albert M Sirunyan et al. Measurement of the $\mathrm{t\bar{t}b\bar{b}}$ production cross section in the all-jet final state in pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Lett. B*, 803:135285, 2020.

[239] Albert M Sirunyan et al. Measurement of the cross section for $\mathrm{t\bar{t}}$ production with additional jets and b jets in pp collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 07:125, 2020.

[240] Georges Aad et al. Measurements of inclusive and differential cross-sections of combined $t\bar{t}\gamma$ and $tW\gamma$ production in the eμ channel at 13 TeV with the ATLAS detector. *JHEP*, 09:049, 2020.

[241] Emanuele R. Nocera, Maria Ubiali, and Cameron Voisey. Single Top Production in PDF fits. *JHEP*, 05:067, 2020.

[242] Georges Aad et al. Comprehensive measurements of t-channel single top-quark production cross sections at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Phys. Rev. D*, 90(11):112006, 2014.

[243] Morad Aaboud et al. Fiducial, total and differential cross-section measurements of t-channel single top-quark production in pp collisions at 8 TeV using data collected by the ATLAS detector. *Eur. Phys. J. C*, 77(8):531, 2017.

[244] Georges Aad et al. Evidence for single top-quark production in the s-channel in proton-proton collisions at $\sqrt{s}$ =8 TeV with the ATLAS detector using the Matrix Element Method. *Phys. Lett. B*, 756:228–246, 2016.

[245] Morad Aaboud et al. Measurement of the inclusive cross-sections of single top-quark and top-antiquark t-channel production in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP*, 04:086, 2017.

[246] Measurement of single top-quark production in the s-channel in proton−proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. 9 2022.

[247] Serguei Chatrchyan et al. Measurement of the Single-Top-Quark t-Channel Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 12:035, 2012.

[248] Vardan Khachatryan et al. Measurement of the t-channel single-top-quark production cross section and of the $| V_{tb} |$ CKM matrix element in pp collisions at $\sqrt{s}$= 8 TeV. *JHEP*, 06:090, 2014.

[249] Vardan Khachatryan et al. Search for s channel single top quark production in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *JHEP*, 09:027, 2016.

[250] Albert M Sirunyan et al. Cross section measurement of t-channel single top quark production in pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Lett. B*, 772:752–776, 2017.

[251] Albert M Sirunyan et al. Measurement of differential cross sections and charge ratios for t-channel single top quark production in proton–proton collisions at $\sqrt{s} = 13$ TeV. *Eur. Phys. J. C*, 80(5):370, 2020.

[252] Georges Aad et al. Measurement of the production cross-section of a single top quark in association with a W boson at 8 TeV with the ATLAS experiment. *JHEP*, 01:064, 2016.

[253] Georges Aad et al. Measurement of single top-quark production in association with a W boson in the single-lepton channel at $\sqrt{s} = 8$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 81(8):720, 2021.

[254] Morad Aaboud et al. Measurement of the cross-section for producing a W boson in association with a single top quark in pp collisions at $\sqrt{s} = 13$ TeV with ATLAS. *JHEP*, 01:063, 2018.

[255] Georges Aad et al. Observation of the associated production of a top quark and a Z boson in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP*, 07:124, 2020.

[256] Serguei Chatrchyan et al. Observation of the associated production of a single top quark and a W boson in pp collisions at $\sqrt{s}$ =8 TeV. *Phys. Rev. Lett.*, 112(23):231802, 2014.

[257] Albert M Sirunyan et al. Measurement of the production cross section for single top quarks in association with W bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 10:117, 2018.

[258] Albert M Sirunyan et al. Observation of Single Top Quark Production in Association with a Z Boson in Proton-Proton Collisions at $\sqrt{s}$ =13 TeV. *Phys. Rev. Lett.*, 122(13):132003, 2019.

[259] Inclusive and differential cross section measurements of single top quark production in association with a Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV. 2021.

[260] Armen Tumasyan et al. Observation of tW production in the single-lepton channel in pp collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 11:111, 2021.

[261] Richard D. Ball, Juan Cruz-Martinez, Luigi Del Debbio, Stefano Forte, Zahari Kassabov, Emanuele R. Nocera, Juan Rojo, Roy Stegeman, and Maria Ubiali. Response to "Parton distributions need representative sampling". 11 2022.

[262] Zahari Kassabov, Emanuele R. Nocera, and Michael Wilson. Regularising experimental correlations in LHC data: theory and application to a global analysis of parton distributions. 7 2022.

[263] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[264] T. Kluge, K. Rabbertz, and M. Wobisch. Fast pQCD calculations for PDF fits. 2006.

[265] M. Wobisch, D. Britzger, T. Kluge, K. Rabbertz, and F. Stober. Theory-Data Comparisons for Jet Measurements in Hadron-Induced Processes. 2011.

[266] Daniel Britzger, Klaus Rabbertz, Fred Stober, and Markus Wobisch. New features in version 2 of the fastNLO project. In *20th International Workshop on Deep-Inelastic Scattering and Related Subjects*, 8 2012.

[267] Valerio Bertone, Rikkert Frederix, Stefano Frixione, Juan Rojo, and Mark Sutton. aMCfast: automation of fast NLO computations for PDF fits. *JHEP*, 08:166, 2014.

[268] Michal Czakon, David Heymes, and Alexander Mitov. Dynamical scales for multi-TeV top-pair production at the LHC. *JHEP*, 04:071, 2017.

[269] Michał Czakon, Zahari Kassabov, Alexander Mitov, Rene Poncelet, and Andrei Popescu. HighTEA: High energy Theory Event Analyser. 4 2023.

[270] Anna Kulesza, Leszek Motyka, Daniel Schwartländer, Tomasz Stebel, and Vincent Theeuwes. Associated production of a top quark pair with a heavy electroweak gauge boson at NLO+NNLL accuracy. *Eur. Phys. J. C*, 79(3):249, 2019.

[271] Edmond L. Berger, Jun Gao, and Hua Xing Zhu. Differential Distributions for t-channel Single Top-Quark Production and Decay at Next-to-Next-to-Leading Order in QCD. *JHEP*, 11:158, 2017.

[272] D. Barducci et al. Interpreting top-quark LHC measurements in the standard-model effective field theory. 2 2018.

[273] Rafael Aoude, Fabio Maltoni, Olivier Mattelaer, Claudio Severi, and Eleni Vryonidou. Renormalisation group effects on SMEFT interpretations of LHC data. 12 2022.

[274] Adam Alloul, Neil D. Christensen, Céline Degrande, Claude Duhr, and Benjamin Fuks. FeynRules 2.0 - A complete toolbox for tree-level phenomenology. *Comput. Phys. Commun.*, 185:2250–2300, 2014.

[275] Céline Degrande, Gauthier Durieux, Fabio Maltoni, Ken Mimasu, Eleni Vryonidou, and Cen Zhang. Automated one-loop computations in the SMEFT. 8 2020.

[276] Richard D. Ball et al. An open-source machine learning framework for global analyses of parton distributions. *Eur. Phys. J. C*, 81(10):958, 2021.

[277] Richard D. Ball et al. A first unbiased global NLO determination of parton distributions and their uncertainties. *Nucl. Phys.*, B838:136, 2010.

[278] Richard D. Ball et al. Parton distributions with LHC data. *Nucl.Phys.*, B867:244, 2013.

[279] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[280] François Chollet et al. Keras. `https://keras.io`, 2015.

[281] J. De Blas et al. `HEPfit`: a code for the combination of indirect and direct constraints on high energy physics models. *Eur. Phys. J. C*, 80(5):456, 2020.

[282] Nuno Castro, Johannes Erdmann, Cornelius Grunwald, Kevin Kröninger, and Nils-Arne Rosien. EFTfitter—A tool for interpreting measurements in the context of effective field theories. *Eur. Phys. J. C*, 76(8):432, 2016.

[283] Ilaria Brivio, Sebastian Bruggisser, Fabio Maltoni, Rhea Moutafis, Tilman Plehn, Eleni Vryonidou, Susanne Westhoff, and C. Zhang. O new physics, where art thou? A global search in the top sector. *JHEP*, 02:131, 2020.

[284] Zhenyu Han and Witold Skiba. Effective theory analysis of precision electroweak data. *Phys. Rev. D*, 71:075009, 2005.

[285] Andy Buckley, Christoph Englert, James Ferrando, David J. Miller, Liam Moore, Michael Russell, and Chris D. White. Constraining top quark effective theory in the LHC Run II era. *JHEP*, 04:015, 2016.

[286] Stefan Bißmann, Johannes Erdmann, Cornelius Grunwald, Gudrun Hiller, and Kevin Kröninger. Constraining top-quark couplings combining top-quark and $\boldsymbol{B}$ decay observables. *Eur. Phys. J. C*, 80(2):136, 2020.

[287] Gauthier Durieux, Adrian Irles, Víctor Miralles, Ana Peñuelas, Roman Pöschl, Martín Perelló, and Marcel Vos. The electro-weak couplings of the top and bottom quarks — Global fit and future prospects. *JHEP*, 12:98, 2019. [Erratum: JHEP 01, 195 (2021)].

[288] Samuel van Beek, Emanuele R. Nocera, Juan Rojo, and Emma Slade. Constraining the SMEFT with Bayesian reweighting. *SciPost Phys.*, 7(5):070, 2019.

[289] Brent R. Yates. Using associated top quark production to probe for new physics within the framework of effective field theory. In *13th International Workshop on Top Quark Physics*, 1 2021.

[290] Johann Brehmer, Kyle Cranmer, Felix Kling, and Tilman Plehn. Better Higgs boson measurements through information geometry. *Phys. Rev. D*, 95(7):073002, 2017.

[291] Rabah Abdul Khalek, Shaun Bailey, Jun Gao, Lucian Harland-Lang, and Juan Rojo. Towards Ultimate Parton Distributions at the High-Luminosity LHC. *Eur. Phys. J. C*, 78(11):962, 2018.

[292] P. Azzi et al. Report from Working Group 1: Standard Model Physics at the HL-LHC and HE-LHC. *CERN Yellow Rep. Monogr.*, 7:1–220, 2019.

[293] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt. Importance Nested Sampling and the MultiNest Algorithm. 2013.

[294] Farhan Feroz and M.P. Hobson. Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis. *Mon. Not. Roy. Astron. Soc.*, 384:449, 2008.

[295] Elie Hammou, Zahari Kassabov, Maeve Madigan, Michelangelo L. Mangano, Luca Mantani, James Moore, Manuel Morales Alvarado, and Maria Ubiali. Hide and seek: how PDFs can conceal new physics. *JHEP*, 11:090, 2023.

[296] Luigi Del Debbio, Tommaso Giani, and Michael Wilson. Bayesian approach to inverse problems: an application to NNPDF closure testing. *Eur. Phys. J. C*, 82(4):330, 2022.

[297] Richard D. Ball et al. The path to proton structure at 1% accuracy. *Eur. Phys. J. C*, 82(5):428, 2022.

[298] Admir Greljo, Shayan Iranipour, Zahari Kassabov, Maeve Madigan, James Moore, Juan Rojo, Maria Ubiali, and Cameron Voisey. Parton distributions in the SMEFT from high-energy Drell-Yan tails. *JHEP*, 07:122, 2021.

[299] James D. Wells and Zhengkang Zhang. Effective theories of universal theories. *JHEP*, 01:123, 2016.

[300] J. de Blas, J. C. Criado, M. Perez-Victoria, and J. Santiago. Effective description of general extensions of the Standard Model: the complete tree-level dictionary. *JHEP*, 03:109, 2018.

[301] B. C. Allanach, Joe Davighi, and Scott Melville. An Anomaly-free Atlas: charting the space of flavour-dependent gauged $U(1)$ extensions of the Standard Model. *JHEP*, 02:082, 2019. [Erratum: JHEP 08, 064 (2019)].

[302] Ennio Salvioni, Giovanni Villadoro, and Fabio Zwirner. Minimal Z-prime models: Present bounds and early LHC reach. *JHEP*, 11:068, 2009.

[303] Ennio Salvioni, Alessandro Strumia, Giovanni Villadoro, and Fabio Zwirner. Non-universal minimal Z' models: present bounds and early LHC reach. *JHEP*, 03:010, 2010.

[304] Paul Langacker. The Physics of Heavy Z' Gauge Bosons. *Rev. Mod. Phys.*, 81:1199–1228, 2009.

[305] Giuliano Panico, Lorenzo Ricci, and Andrea Wulzer. High-energy EFT probes with fully differential Drell-Yan measurements. 3 2021.

[306] Riccardo Torre, Lorenzo Ricci, and Andrea Wulzer. On the W&Y interpretation of high-energy Drell-Yan measurements. *JHEP*, 02:144, 2021.

[307] R. L. Workman et al. Review of Particle Physics. *PTEP*, 2022:083C01, 2022.

[308] Radja Boughezal, Emanuele Mereghetti, and Frank Petriello. Dilepton production in the SMEFT at O(1/Λ4). *Phys. Rev. D*, 104(9):095022, 2021.

[309] Georges Aad et al. Measurement of the high-mass Drell–Yan differential cross-section in pp collisions at sqrt(s)=7 TeV with the ATLAS detector. *Phys. Lett. B*, 725:223–242, 2013.

[310] Georges Aad et al. Measurement of the double-differential high-mass Drell-Yan cross section in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 08:009, 2016.

[311] Serguei Chatrchyan et al. Measurement of the Differential and Double-Differential Drell-Yan Cross Sections in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 12:030, 2013.

[312] Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector. 4 2020.

[313] Morad Aaboud et al. Measurement of $W^{\pm}Z$ production cross sections and gauge boson polarisation in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 79(6):535, 2019.

[314] Albert M Sirunyan et al. Measurements of the pp → WZ inclusive and differential production cross section and constraints on charged anomalous triple gauge couplings at $\sqrt{s} = 13$ TeV. *JHEP*, 04:122, 2019.

[315] Stefano Carrazza, Stefano Forte, Zahari Kassabov, and Juan Rojo. Specialized minimal PDFs for optimized LHC calculations. *Eur. Phys. J. C*, 76(4):205, 2016.

[316] R. S. Towell et al. Improved measurement of the anti-d / anti-u asymmetry in the nucleon sea. *Phys. Rev. D*, 64:052002, 2001.

[317] Victor Mukhamedovich Abazov et al. Measurement of the Muon Charge Asymmetry in $p\bar{p} \to$ W+X $\to \mu\nu$ + X Events at $\sqrt{s}$=1.96 TeV. *Phys. Rev. D*, 88:091102, 2013.

[318] Morad Aaboud et al. Precision measurement and interpretation of inclusive W^+ , W^- and Z/γ^* production cross sections with the ATLAS detector. *Eur. Phys. J. C*, 77(6):367, 2017.

Appendix A

Transformation to standard DIS variables

In Chapter 1, we changed variables from the three-momentum of the outgoing electron $\mathbf{p}_{\ell'}$ to the standard DIS variables (x, y, ϕ). To make this transformation, we note that in $d = 4$, we have:

$$d^3\mathbf{p}_{\ell'} = |\mathbf{p}_{\ell'}|^2 d|\mathbf{p}_{\ell'}| d\cos(\theta) d\phi = |\mathbf{p}_{\ell'}|^2 \left|\det\left(\frac{\partial(|\mathbf{p}_{\ell'}|, \cos(\theta))}{\partial(x, y)}\right)\right| dx dy d\phi, \tag{A.1}$$

where θ is the angle defined by $\cos(\theta) := \mathbf{p}_\ell \cdot \mathbf{p}_{\ell'}/|\mathbf{p}_\ell||\mathbf{p}_{\ell'}|$, and $d\phi$ is the remaining angular integral. To calculate the Jacobian matrix in Eq. (A.2), consider working in the rest frame of the proton, so that $\mathbf{p}_H = \mathbf{0}$. Then, assuming the masslessness of the electrons so that $E_\ell = |\mathbf{p}_\ell|$, $E_{\ell'} = |\mathbf{p}_{\ell'}|$, we can rewrite the variables in Eq. (1.15) through:

$$x = \frac{|\mathbf{p}_\ell||\mathbf{p}_{\ell'}|\,(1 - \cos(\theta))}{M_H(|\mathbf{p}_\ell| - |\mathbf{p}_{\ell'}|)}, \qquad y = 1 - \frac{|\mathbf{p}_{\ell'}|}{|\mathbf{p}_\ell|}, \tag{A.2}$$

where M_H is the rest mass of the proton. Solving these equations simultaneously for $|\mathbf{p}_{\ell'}|$ and $\cos(\theta)$, we have:

$$|\mathbf{p}_{\ell'}| = |\mathbf{p}_\ell|(1 - y), \qquad \cos(\theta) = 1 - \frac{M_H x y}{|\mathbf{p}_\ell|(1 - y)}. \tag{A.3}$$

This allows us to compute the Jacobian factor:

$$\left|\det\left(\frac{\partial(|\mathbf{p}_{\ell'}|, \cos(\theta))}{\partial(x, y)}\right)\right| = \frac{M_H y}{1 - y}. \tag{A.4}$$

Hence the measure in Eq. (A.2) may be rewritten as:

$$d^3\mathbf{p}_{\ell'} = |\mathbf{p}_\ell|^2(1 - y) M_H y \; dx dy d\phi. \tag{A.5}$$

Appendix B

Proofs of plus prescription identities

In this Appendix, we give a proof of the plus prescription identities used in the introductory chapter.

Theorem 1. The following distributional identity holds:

$$\frac{u^\epsilon}{(1-u)^{\epsilon+1}} = -\frac{1}{\epsilon}\delta(1-u) + \left(\frac{1}{1-u}\right)_+ - \epsilon\left(\frac{\log(1-u)}{1-u}\right)_+ + \epsilon\frac{\log(u)}{1-u} + O(\epsilon^2). \tag{B.1}$$

Proof: Let f be an arbitrary smooth function. Then:

$$\int_0^1 \frac{f(u)}{(1-u)^{\epsilon+1}}\, du = \int_0^1 \frac{(f(u)-f(1))}{1-u}\cdot(1-u)^{-\epsilon}\, du + f(1)\int_0^1 \frac{1}{(1-u)^{\epsilon+1}}\, du \tag{B.2}$$

$$= \int_0^1 \frac{(f(u)-f(1))}{1-u}(1-\epsilon\log(1-u))\, du + f(1)\left[\frac{(1-u)^\epsilon}{\epsilon}\right]_0^1 + O(\epsilon^2) \tag{B.3}$$

$$= \int_0^1 \left[\left(\frac{1}{1-u}\right)_+ f(u) - \epsilon\left(\frac{\log(1-u)}{1-u}\right)_+ f(u)\right] du - \frac{f(1)}{\epsilon} + O(\epsilon^2) \tag{B.4}$$

$$= \int_0^1 \left[\left(\frac{1}{1-u}\right)_+ f(u) - \epsilon\left(\frac{\log(1-u)}{1-u}\right)_+ f(u) - \frac{\delta(1-u)}{\epsilon}f(u)\right] du + O(\epsilon^2). \tag{B.5}$$

Hence, as distributions, we have:

$$\frac{1}{(1-u)^{\epsilon+1}} = -\frac{1}{\epsilon}\delta(1-u) + \left(\frac{1}{1-u}\right)_+ - \epsilon\left(\frac{\log(1-u)}{1-u}\right)_+ + O(\epsilon^2). \tag{B.6}$$

Multiplying both sides by $u^\epsilon = 1 + \epsilon \log(u) + O(\epsilon^2)$, we find that:

$$\frac{u^\epsilon}{(1-u)^{\epsilon+1}} = -\frac{1}{\epsilon}\delta(1-u) + \left(\frac{1}{1-u}\right)_+ + \epsilon\left(\frac{1}{1-u}\right)_+ \log(u) - \epsilon\left(\frac{\log(1-u)}{1-u}\right)_+ + O(\epsilon^2). \tag{B.7}$$

But the plus prescription multiplied by $\log(u)$ has no effect, so we can just remove the plus. Therefore, we're done. □

Theorem 2. The following distributional identity holds:

$$\left(-\frac{9}{2} - \frac{\pi^2}{3}\right)\delta(1-u) + 3 + 2u - \frac{3}{2(1-u)_+} + (1+u^2)\left(\frac{\log(1-u)}{1-u}\right)_+ - \left(\frac{1+u^2}{1-u}\right)\log(u)$$
$$= \left[\frac{1+u^2}{1-u}\left(\log\left(\frac{1-u}{u}\right) - \frac{3}{4}\right) + \frac{5u+9}{4}\right]_+. \tag{B.8}$$

Proof: We work from the right hand side to the left hand side. Let f be an arbitrary smooth function. Then:

$$\int_0^1 du\ \left[\frac{1+u^2}{1-u}\left(\log\left(\frac{1-u}{u}\right) - \frac{3}{4}\right) + \frac{5u+9}{4}\right]_+ f(u)$$
$$= \int_0^1 du\ \left[\frac{1+u^2}{1-u}\left(\log(1-u) - \log(u) - \frac{3}{4}\right) + \frac{5u+9}{4}\right](f(u) - f(1)) \tag{B.9}$$
$$= \int_0^1 du\ \left[\frac{\log(1-u)}{1-u}((1+u^2)f(u) - 2f(1)) + (1+u)\log(1-u)f(1)\right.$$
$$\left. - \frac{1+u^2}{1-u}\log(u)f(u) + \frac{1+u^2}{1-u}\log(u)f(1)\right] + \left[-\frac{3}{4}\left(\frac{1+u^2}{1-u}\right) + \frac{5u+9}{4}\right](f(u) - f(1)) \tag{B.10}$$

Now observe that we can perform some of the integrals:

$$\int_0^1 du\ (1+u)\log(1-u) = -\frac{7}{4}, \qquad \int_0^1 du\ \frac{1+u^2}{1-u}\log(u) = \frac{5}{4} - \frac{\pi^2}{3}. \tag{B.11}$$

We can also perform some manipulation of the non-logarithmic terms:

$$\left[-\frac{3}{4}\left(\frac{1+u^2}{1-u}\right)+\frac{5u+9}{4}\right](f(u)-f(1))$$
$$=(3+2u)f(u)-\frac{3}{2(1-u)}(f(u)-f(1))-\left(\frac{3}{4}(1+u)+\frac{5u+9}{4}\right)f(1). \qquad \text{(B.12)}$$

Performing all the integrals adjacent to $f(1)$, we obtain the required left hand side. $\square$

Appendix C

Random seed dependence

As described in Eq. (5.1), the pseudodata used in Chapter 5 is stochastic, fluctuated around the supposed law of Nature in order to simulate random experimental noise. This noise is generated in a reproducible manner using the NNPDF closure test code by selecting a particular *seed* for the generation algorithm; different choices of seed lead to different choices of noise.

This has consequences for the resulting contaminated PDF fits, which in principle can depend on the seed used for the random noise. In certain parts of this work, in particular in the production of Figs. 5.5 and 5.7, we have made the approximation that the contaminated PDFs do not depend significantly on the choice of random seed; rather, we hope that their behaviour is most importantly affected by whether or not New Physics is present in the pseudodata or not. This is a useful approximation to make, since it avoids the requirement of running a large quantity of PDF fits, which is computationally expensive.

We justify this approximation in this brief appendix by comparing the PDF luminosities in various contaminated fits produced using different seeds for the random pseudodata. The luminosities are the relevant quantity to compare, since these are the quantities which enter the theoretical predictions for the hadronic data, in particular the Drell-Yan data, the focus of this study.

In Fig. C.1, we plot the luminosities obtained from contaminated fits resulting from setting the $\hat{W}$ parameter to the benchmark values $\hat{W} = 3 \times 10^{-5}$, $\hat{W} = 8 \times 10^{-5}$ and $\hat{W} = 15 \times 10^{-5}$. We display the results for two separate contaminated fits for each of the benchmark values; in each case, one of the fits results from the use of a particular random seed (called *seed* 1 in the plots), whilst the other results from the use of another random seed (called *seed* 2 in the plots). We observe that the luminosities are completely statistically equivalent between the two seeds, but that across different benchmark values of $\hat{W}$, there is indeed a statistical difference between the luminosities. This justifies that the leading effect on the contaminated fits is the injection of New Physics into the pseudodata,

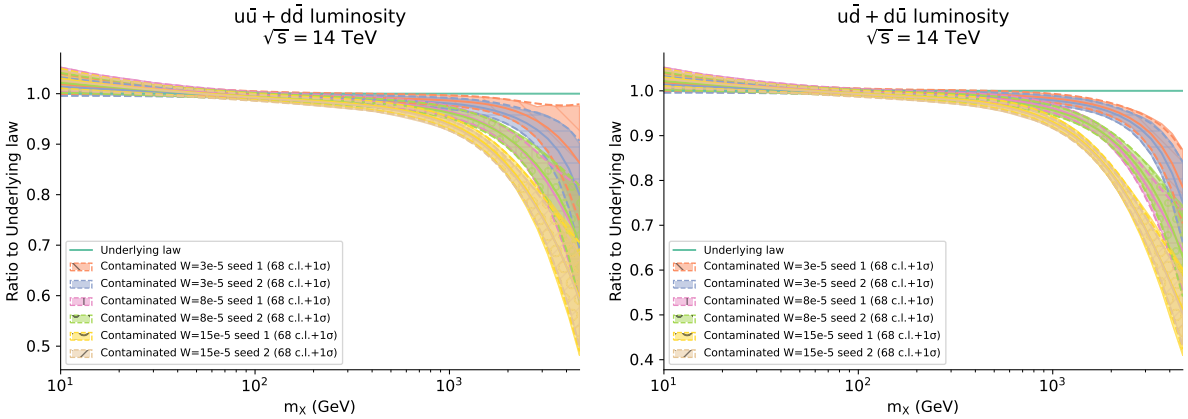


Figure C.1: Comparison between luminosities obtained in contaminated fits using two different random seeds in the generation of pseudodata. In each case, we display six contaminated fits: two fits for each of the benchmark values $\hat{W} = 3 \times 10^{-5}, 8 \times 10^{-5}, 15 \times 10^{-5}$, trained on pseudodata generated with random seed 1 and random seed 2 respectively.

rather than the random noise added to the pseudodata. In particular, the approximation in Sect. 5.3 is fully justified. Similar conclusions hold for the $\hat{Y}$ parameter.

Appendix D

Contaminated fit quality

In this appendix we give details about the fit-quality of the closure tests presented in Chapter 5. In Table D.1 and D.2, we list the value of the reduced $\chi^2/n_{\rm dat}$ as well as of the n_σ estimator (see Sect. 5.1.3 for details) for each dataset included in the fit, under all the contamination scenarios we have tested. We have highlighted the datasets whose fit quality deteriorates the most in Figs. D.1 and D.2. In particular, the two figures showcase the tension between the fixed-target datasets and the HL-LHC projected data as the value of the $\hat{W}$ increases.

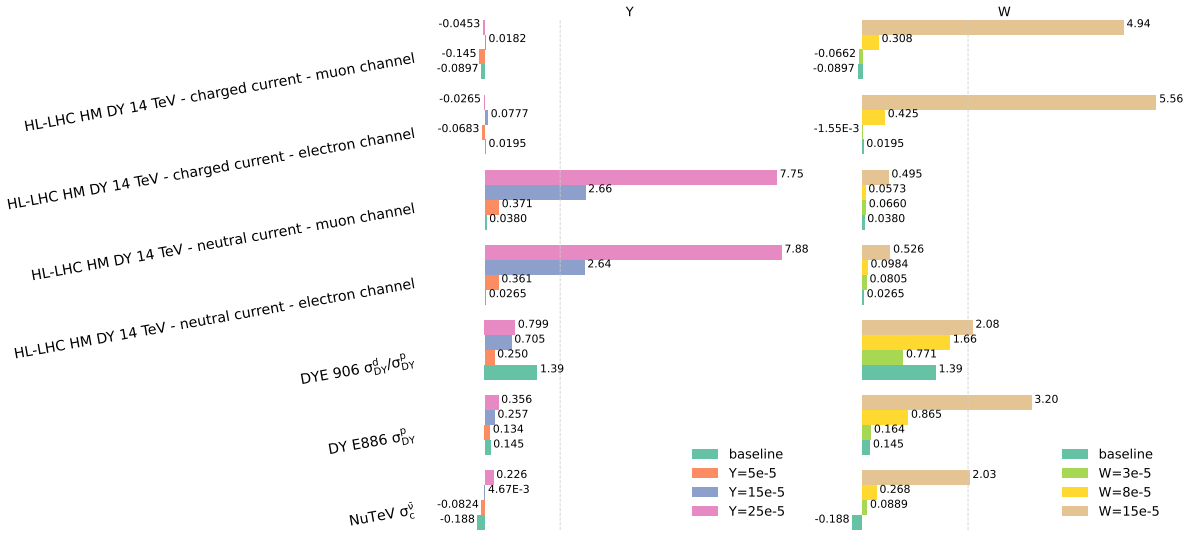


Figure D.1: Value of n_σ, defined in Eq. (5.4) for all datasets that pass the threshold criterion of $n_\sigma > 2$ discussed in Sect. 5.1.3 in each of the three fits performed by injecting various degrees of New Physics. The figure on the left, New Physics signals in the data are added according to Scenario I (flavour-universal Z' model), namely the baseline $\hat{Y} = 0$ (green bars), $\hat{Y} = 5 \cdot 10^{-5}$ (orange bars), $\hat{Y} = 15 \cdot 10^{-5}$ (blue bars) and $\hat{Y} = 25 \cdot 10^{-5}$ (pink bars). In the figure on the right, signals are added according to Scenario II (flavour-universal W' model), namely the baseline $\hat{W} = 0$ (again, green bars), $\hat{W} = 3 \cdot 10^{-5}$ (light green bars), $\hat{W} = 8 \cdot 10^{-5}$ (yellow bars) and $\hat{W} = 15 \cdot 10^{-5}$ (brown bars)

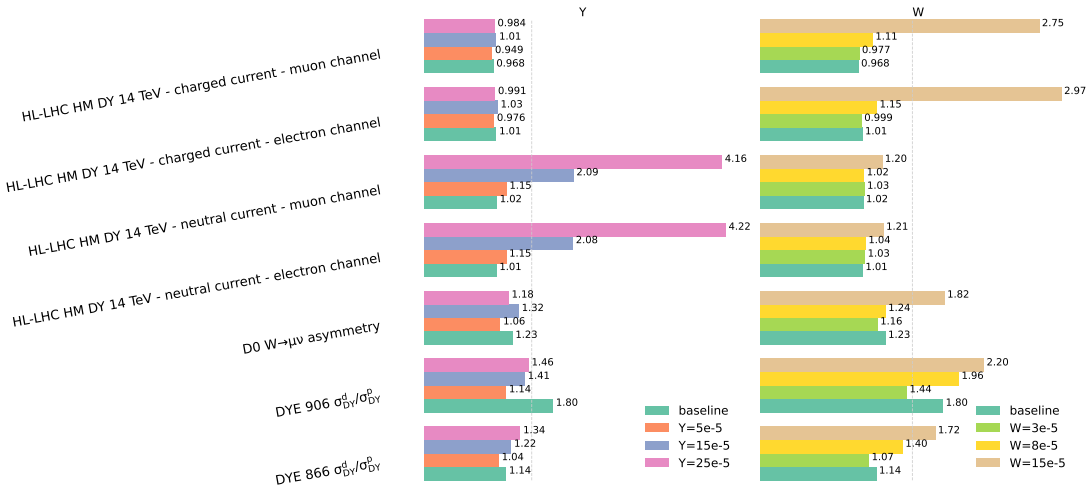


Figure D.2: $\chi^2/n_{\rm dat}$ distribution for all datasets that pass the threshold criterion of $\chi^2/n_{\rm dat} > 1.5$ discussed in Sect. 5.1.3 in each of the three fits performed by injecting various degrees of New Physics signals in the data according to Scenario I (left panel) and Scenario II (right panel)

	baseline		Y=5e-5		Y=15e-5		Y=25e-5	
	χ^2	n_σ	χ^2	n_σ	χ^2	n_σ	χ^2	n_σ
NMC d/p	1.02	0.14	1.02	0.12	1.03	0.24	1.06	0.45
NMC p	1.03	0.26	1.02	0.23	1.02	0.18	1.02	0.18
SLAC p	1.02	0.06	1.01	0.05	1.01	0.03	1.02	0.07
SLAC d	1.00	-0.01	0.98	-0.07	0.99	-0.02	0.99	-0.04
BCDMS p	1.02	0.20	1.00	0.06	1.02	0.21	1.01	0.11
BCDMS d	1.01	0.07	1.00	0.01	1.01	0.08	1.00	0.03
CHORUS σ^ν_{CC}	1.00	0.02	1.00	-0.06	1.00	0.06	1.00	0.01
CHORUS σ^ν_{CC}	0.99	-0.13	0.99	-0.13	1.00	-0.03	0.99	-0.08
NuTeV σ^ν_c	0.99	-0.06	0.99	-0.05	0.99	-0.02	1.00	0.02
NuTeV $\sigma^{\bar{\nu}}_c$	0.96	-0.19	0.98	-0.08	1.00	0.00	1.05	0.23
HERA I+II inclusive NC e^-p	1.00	-0.02	1.01	0.12	1.00	0.02	1.02	0.17
HERA I+II inclusive NC e^+p 460 GeV	1.01	0.08	1.01	0.12	1.01	0.13	1.02	0.18
HERA I+II inclusive NC e^+p 575 GeV	0.98	-0.21	1.00	0.01	0.99	-0.17	1.01	0.07
HERA I+II inclusive NC e^+p 820 GeV	1.00	-0.00	1.01	0.07	1.00	-0.01	1.01	0.09
HERA I+II inclusive NC e^+p 920 GeV	1.02	0.29	1.05	0.63	1.03	0.35	1.05	0.67
HERA I+II inclusive CC e^-p	0.99	-0.05	1.03	0.13	0.99	-0.03	1.03	0.15
HERA I+II inclusive CC e^+p	1.02	0.08	1.02	0.07	1.03	0.11	1.04	0.18
HERA comb. $\sigma^{\rm red}_{c\bar{c}}$	1.00	0.02	1.02	0.09	1.01	0.05	1.03	0.11
HERA comb. $\sigma^{\rm red}_{b\bar{b}}$	1.12	0.43	1.13	0.45	1.14	0.49	1.13	0.48
DYE 866 $\sigma^d_{\rm DY}/\sigma^p_{\rm DY}$	1.14	0.40	1.04	0.12	1.22	0.59	1.34	0.94
DY E886 $\sigma^p_{\rm DY}$	1.02	0.14	1.02	0.13	1.04	0.26	1.05	0.36
DY E605 $\sigma^p_{\rm DY}$	1.08	0.53	1.07	0.43	1.06	0.42	1.06	0.39
DYE 906 $\sigma^d_{\rm DY}/\sigma^p_{\rm DY}$	1.80	1.39	1.14	0.25	1.41	0.70	1.46	0.80
CDF Z rapidity (new)	1.06	0.21	1.03	0.12	1.06	0.21	1.01	0.06
D0 Z rapidity	1.03	0.10	1.02	0.08	1.04	0.17	1.03	0.11
D0 $W \to \mu\nu$ asymmetry	1.23	0.50	1.06	0.13	1.32	0.69	1.18	0.38
ATLAS W, Z 7 TeV 2010	1.05	0.20	1.04	0.17	1.05	0.20	1.05	0.20
ATLAS HM DY 7 TeV	1.02	0.04	1.05	0.12	1.01	0.02	1.03	0.09
ATLAS low-mass DY 2011	0.90	-0.17	1.04	0.07	0.87	-0.22	1.01	0.01
ATLAS W, Z 7 TeV 2011 Central selection	1.06	0.28	1.07	0.36	1.07	0.31	1.07	0.35
ATLAS W, Z 7 TeV 2011 Forward selection	0.91	-0.25	1.33	0.90	0.90	-0.29	1.32	0.87
ATLAS DY 2D 8 TeV high mass	1.02	0.11	1.03	0.13	1.02	0.08	1.03	0.12
ATLAS DY 2D 8 TeV low mass	1.03	0.16	1.00	0.00	1.02	0.13	0.99	-0.04
ATLAS W, Z inclusive 13 TeV	1.07	0.09	1.08	0.10	1.09	0.11	1.13	0.16
ATLAS W^++jet 8 TeV	1.17	0.46	0.96	-0.11	1.17	0.48	0.95	-0.13
ATLAS W^-+jet 8 TeV	1.19	0.51	0.97	-0.09	1.20	0.56	0.97	-0.08
ATLAS Z p_T 8 TeV (p^{ll}_T, M_{ll})	1.01	0.03	0.98	-0.07	1.01	0.04	0.99	-0.05
ATLAS Z p_T 8 TeV (p^{ll}_T, y_{ll})	0.98	-0.10	0.94	-0.31	0.98	-0.10	0.93	-0.36
ATLAS $\sigma^{\rm tot}_{tt}$	1.03	0.02	1.14	0.10	1.05	0.03	1.19	0.14
ATLAS $\sigma^{\rm tot}_{tt}$ 8 TeV	1.31	0.22	1.12	0.08	1.27	0.19	1.08	0.06
ATLAS $\sigma^{\rm tot}_{tt}$ 13 TeV Run II full lumi	0.92	-0.06	0.93	-0.05	0.97	-0.02	0.98	-0.01
ATLAS $t\bar{t}$ y_t	1.03	0.05	1.06	0.08	1.04	0.06	1.04	0.05
ATLAS $t\bar{t}$ $y_{t\bar{t}}$	1.04	0.05	1.04	0.05	1.06	0.08	1.05	0.07
ATLAS $t\bar{t}$ normalised $\lvert y_t \rvert$	1.13	0.21	1.13	0.20	1.15	0.24	1.17	0.26
ATLAS jets 8 TeV, R=0.6	0.83	-1.53	0.94	-0.58	0.83	-1.55	0.94	-0.60
ATLAS dijets 7 TeV, R=0.6	1.03	0.19	1.00	-0.00	1.04	0.24	1.01	0.09
ATLAS direct photon production 13 TeV	0.97	-0.16	1.03	0.14	0.98	-0.11	1.04	0.21
ATLAS single top R_t 7 TeV	1.14	0.10	1.26	0.18	1.05	0.03	1.15	0.11
ATLAS single top R_t 13 TeV	0.91	-0.07	1.01	0.01	0.93	-0.05	1.04	0.03
ATLAS single top y_t (normalised)	0.94	-0.07	1.07	0.09	0.93	-0.09	1.04	0.04
ATLAS single antitop y (normalised)	0.92	-0.10	0.91	-0.11	0.97	-0.04	0.98	-0.03
CMS W asymmetry 840 pb	0.99	-0.03	0.98	-0.04	0.98	-0.04	1.00	-0.01
CMS W asymmetry 4.7 fb	0.97	-0.07	0.97	-0.06	0.97	-0.08	1.00	0.00
CMS Drell-Yan 2D 7 TeV 2011	1.01	0.05	1.01	0.07	1.00	0.04	1.01	0.10
CMS W rapidity 8 TeV	1.06	0.21	1.12	0.39	1.07	0.22	1.13	0.42
CMS Z p_T 8 TeV (p^{ll}_T, y_{ll})	1.03	0.12	1.03	0.11	1.03	0.12	1.04	0.14
CMS dijets 7 TeV	0.97	-0.15	1.05	0.24	0.97	-0.14	1.05	0.25
CMS jets 8 TeV	0.99	-0.11	1.00	-0.03	0.99	-0.09	1.00	-0.01
CMS $\sigma^{\rm tot}_{tt}$ 7 TeV	0.86	-0.10	0.95	-0.03	0.86	-0.10	1.00	0.00
CMS $\sigma^{\rm tot}_{tt}$ 8 TeV	1.18	0.13	1.09	0.06	1.21	0.15	1.07	0.05
CMS $\sigma^{\rm tot}_{tt}$ 13 TeV	0.98	-0.01	1.11	0.08	0.99	-0.00	1.12	0.09
CMS $t\bar{t}$ rapidity $y_{t\bar{t}}$	1.06	0.12	1.04	0.08	1.04	0.09	1.01	0.02
CMS $\sigma^{\rm tot}_{tt}$ 5 TeV	0.86	-0.10	0.77	-0.17	0.82	-0.13	0.75	-0.18
CMS $t\bar{t}$ double differential $(m_{t\bar{t}}, y_{t\bar{t}})$	0.99	-0.04	1.00	0.01	1.02	0.04	1.03	0.07
CMS $t\bar{t}$ absolute y_t	1.01	0.03	1.02	0.04	1.02	0.05	1.03	0.06
CMS $t\bar{t}$ absolute $\lvert y_t \rvert$	0.98	-0.05	0.99	-0.03	0.97	-0.06	0.96	-0.09
CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV	0.93	-0.05	0.91	-0.06	0.89	-0.08	0.86	-0.10
CMS single top R_t 8 TeV	0.64	-0.26	1.21	0.15	0.63	-0.26	1.14	0.10
CMS single top R_t 13 TeV	1.50	0.35	1.44	0.31	1.46	0.33	1.42	0.30
LHCb Z 940 pb	1.08	0.17	0.96	-0.09	1.11	0.24	0.97	-0.06
LHCb $Z \to ee$ 2 fb	1.03	0.08	1.04	0.13	1.01	0.02	1.04	0.11
LHCb $W, Z \to \mu$ 7 TeV	0.98	-0.07	0.97	-0.10	1.02	0.06	1.01	0.05
LHCb $W, Z \to \mu$ 8 TeV	1.08	0.32	1.13	0.51	1.09	0.35	1.18	0.68
LHCb $Z \to \mu\mu$	1.09	0.26	1.05	0.15	1.10	0.27	1.05	0.13
LHCb $Z \to ee$	1.07	0.19	1.03	0.08	1.07	0.19	1.04	0.10
CMS HM DY 8 TeV	0.98	-0.09	0.98	-0.08	0.97	-0.13	0.98	-0.10
CMS HM DY 13 TeV - combined channel	0.97	-0.14	0.97	-0.16	0.97	-0.14	0.97	-0.15
HL-LHC HM DY 14 TeV - neutral current - electron channel	1.01	0.03	1.15	0.36	2.08	2.64	4.22	7.88
HL-LHC HM DY 14 TeV - neutral current - muon channel	1.02	0.04	1.15	0.37	2.09	2.66	4.16	7.75
HL-LHC HM DY 14 TeV - charged current - electron channel	1.01	0.02	0.98	-0.07	1.03	0.08	0.99	-0.03
HL-LHC HM DY 14 TeV - charged current - muon channel	0.97	-0.09	0.95	-0.14	1.01	0.02	0.98	-0.05

Table D.1: Fit quality in fits contaminated with the Y operator.

	baseline		W=3e-5		W=8e-5		W=15e-5	
	χ^2	n_σ	χ^2	n_σ	χ^2	n_σ	χ^2	n_σ
NMC d/p	1.02	0.14	1.01	0.04	1.04	0.31	1.05	0.42
NMC p	1.03	0.26	1.03	0.27	1.02	0.22	1.03	0.28
SLAC p	1.02	0.06	1.02	0.07	1.01	0.03	1.02	0.06
SLAC d	1.00	-0.01	0.98	-0.07	0.99	-0.05	1.00	0.02
BCDMS p	1.02	0.20	1.01	0.07	1.02	0.24	1.01	0.11
BCDMS d	1.01	0.07	1.00	0.01	1.01	0.10	1.00	0.02
CHORUS σ^{ν}_{CC}	1.00	0.02	1.00	-0.07	1.00	0.06	1.00	0.04
CHORUS $\sigma^{\bar{\nu}}_{CC}$	0.99	-0.13	0.99	-0.13	1.00	-0.00	1.00	-0.02
NuTeV σ^{ν}_{c}	0.99	-0.06	0.99	-0.05	1.01	0.05	1.00	0.01
NuTeV $\sigma^{\bar{\nu}}_{c}$	0.96	-0.19	1.02	0.09	1.06	0.27	1.47	2.03
HERA I+II inclusive NC e^-p	1.00	-0.02	1.01	0.13	1.00	0.03	1.02	0.19
HERA I+II inclusive NC e^+p 460 GeV	1.01	0.08	1.01	0.12	1.01	0.12	1.02	0.20
HERA I+II inclusive NC e^+p 575 GeV	0.98	-0.21	1.00	0.01	0.98	-0.18	1.01	0.10
HERA I+II inclusive NC e^+p 820 GeV	1.00	-0.00	1.01	0.07	1.00	-0.02	1.02	0.10
HERA I+II inclusive NC e^+p 920 GeV	1.02	0.29	1.06	0.76	1.04	0.54	1.09	1.23
HERA I+II inclusive CC e^-p	0.99	-0.05	1.03	0.13	1.00	-0.00	1.03	0.15
HERA I+II inclusive CC e^+p	1.02	0.08	1.02	0.08	1.04	0.19	1.10	0.45
HERA comb. $\sigma^{\rm red}_{c\bar{c}}$	1.00	0.02	1.02	0.08	1.01	0.02	1.01	0.04
HERA comb. $\sigma^{\rm red}_{b\bar{b}}$	1.12	0.43	1.13	0.45	1.13	0.48	1.13	0.47
DYE 866 $\sigma^{d}_{\rm DY}/\sigma^{p}_{\rm DY}$	1.14	0.40	1.07	0.20	1.40	1.11	1.72	1.98
DY E886 $\sigma^{p}_{\rm DY}$	1.02	0.14	1.02	0.16	1.13	0.87	1.48	3.20
DY E605 $\sigma^{p}_{\rm DY}$	1.08	0.53	1.07	0.44	1.07	0.47	1.08	0.50
DYE 906 $\sigma^{d}_{\rm DY}/\sigma^{p}_{\rm DY}$	1.80	1.39	1.44	0.77	1.96	1.66	2.20	2.08
CDF Z rapidity (new)	1.06	0.21	1.03	0.12	1.06	0.22	1.02	0.07
D0 Z rapidity	1.03	0.10	1.02	0.07	1.04	0.16	1.02	0.07
D0 $W \to \mu\nu$ asymmetry	1.23	0.50	1.16	0.33	1.24	0.50	1.82	1.73
ATLAS W, Z 7 TeV 2010	1.05	0.20	1.04	0.17	1.06	0.22	1.05	0.18
ATLAS HM DY 7 TeV	1.02	0.04	1.05	0.12	1.01	0.04	1.03	0.06
ATLAS low-mass DY 2011	0.90	-0.17	1.04	0.07	0.87	-0.23	0.99	-0.02
ATLAS W, Z 7 TeV 2011 Central selection	1.06	0.28	1.07	0.35	1.06	0.28	1.08	0.37
ATLAS W, Z 7 TeV 2011 Forward selection	0.91	-0.25	1.33	0.90	0.90	-0.29	1.31	0.84
ATLAS DY 2D 8 TeV high mass	1.02	0.11	1.03	0.14	1.02	0.10	1.04	0.20
ATLAS DY 2D 8 TeV low mass	1.03	0.16	1.00	0.00	1.03	0.16	0.99	-0.04
ATLAS W, Z inclusive 13 TeV	1.07	0.09	1.07	0.09	1.09	0.11	1.08	0.10
ATLAS W^++jet 8 TeV	1.17	0.46	0.96	-0.10	1.17	0.48	0.96	-0.12
ATLAS W^-+jet 8 TeV	1.19	0.51	0.97	-0.10	1.21	0.58	0.98	-0.06
ATLAS Z p_T 8 TeV (p_T^{ll}, M_{ll})	1.01	0.03	0.98	-0.07	1.01	0.03	0.99	-0.05
ATLAS Z p_T 8 TeV (p_T^{ll}, y_{ll})	0.98	-0.10	0.94	-0.29	0.99	-0.06	0.96	-0.21
ATLAS $\sigma^{\rm tot}_{tt}$	1.03	0.02	1.14	0.10	1.04	0.03	1.17	0.12
ATLAS $\sigma^{\rm tot}_{tt}$ 8 TeV	1.31	0.22	1.12	0.09	1.30	0.21	1.13	0.09
ATLAS $\sigma^{\rm tot}_{tt}$ 13 TeV Run II full lumi	0.92	-0.06	0.93	-0.05	0.93	-0.05	0.97	-0.02
ATLAS $t\bar{t}$ y_t	1.03	0.05	1.06	0.09	1.03	0.04	1.06	0.08
ATLAS $t\bar{t}$ $y_{t\bar{t}}$	1.04	0.05	1.04	0.06	1.05	0.08	1.09	0.12
ATLAS $t\bar{t}$ normalised $\lvert y_t \rvert$	1.13	0.21	1.13	0.21	1.14	0.22	1.18	0.28
ATLAS jets 8 TeV, R=0.6	0.83	-1.53	0.94	-0.57	0.83	-1.53	0.94	-0.54
ATLAS dijets 7 TeV, R=0.6	1.03	0.19	1.00	0.00	1.03	0.18	1.01	0.10
ATLAS direct photon production 13 TeV	0.97	-0.16	1.03	0.14	0.98	-0.13	1.03	0.16
ATLAS single top R_t 7 TeV	1.14	0.10	1.25	0.18	1.06	0.04	1.16	0.11
ATLAS single top R_t 13 TeV	0.91	-0.07	1.01	0.01	0.94	-0.04	1.05	0.03
ATLAS single top y_t (normalised)	0.94	-0.07	1.07	0.09	0.94	-0.08	1.04	0.04
ATLAS single antitop y (normalised)	0.92	-0.10	0.91	-0.11	0.94	-0.07	0.98	-0.03
CMS W asymmetry 840 pb	0.99	-0.03	0.99	-0.02	0.97	-0.08	1.05	0.12
CMS W asymmetry 4.7 fb	0.97	-0.07	0.97	-0.06	0.97	-0.06	0.97	-0.06
CMS Drell-Yan 2D 7 TeV 2011	1.01	0.05	1.01	0.07	1.01	0.04	1.01	0.08
CMS W rapidity 8 TeV	1.06	0.21	1.11	0.38	1.07	0.25	1.11	0.38
CMS Z p_T 8 TeV (p_T^{ll}, y_{ll})	1.03	0.12	1.03	0.12	1.04	0.13	1.06	0.21
CMS dijets 7 TeV	0.97	-0.15	1.05	0.24	0.97	-0.13	1.05	0.28
CMS jets 8 TeV	0.99	-0.11	1.00	-0.02	0.99	-0.05	1.01	0.06
CMS $\sigma^{\rm tot}_{tt}$ 7 TeV	0.86	-0.10	0.95	-0.03	0.86	-0.10	0.99	-0.00
CMS $\sigma^{\rm tot}_{tt}$ 8 TeV	1.18	0.13	1.08	0.06	1.22	0.16	1.07	0.05
CMS $\sigma^{\rm tot}_{tt}$ 13 TeV	0.98	-0.01	1.11	0.08	0.99	-0.01	1.13	0.09
CMS $t\bar{t}$ rapidity $y_{t\bar{t}}$	1.06	0.12	1.04	0.08	1.03	0.07	1.02	0.05
CMS $\sigma^{\rm tot}_{tt}$ 5 TeV	0.86	-0.10	0.77	-0.16	0.81	-0.13	0.73	-0.19
CMS $t\bar{t}$ double differential $(m_{t\bar{t}}, y_{t\bar{t}})$	0.99	-0.04	1.01	0.02	1.01	0.04	1.03	0.08
CMS $t\bar{t}$ absolute y_t	1.01	0.03	1.02	0.04	1.02	0.04	1.05	0.11
CMS $t\bar{t}$ absolute $\lvert y_t \rvert$	0.98	-0.05	0.99	-0.03	0.98	-0.05	0.96	-0.10
CMS single top $\sigma_t + \sigma_{\bar{t}}$ 7 TeV	0.93	-0.05	0.91	-0.06	0.88	-0.09	0.86	-0.10
CMS single top R_t 8 TeV	0.64	-0.26	1.21	0.15	0.65	-0.25	1.15	0.10
CMS single top R_t 13 TeV	1.50	0.35	1.44	0.31	1.46	0.32	1.40	0.28
LHCb Z 940 pb	1.08	0.17	0.95	-0.10	1.12	0.25	0.96	-0.08
LHCb $Z \to ee$ 2 fb	1.03	0.08	1.04	0.12	1.01	0.02	1.01	0.03
LHCb $W, Z \to \mu$ 7 TeV	0.98	-0.07	0.96	-0.17	1.07	0.26	1.13	0.48
LHCb $W, Z \to \mu$ 8 TeV	1.08	0.32	1.12	0.45	1.17	0.65	1.32	1.22
LHCb $Z \to \mu\mu$	1.09	0.26	1.05	0.14	1.10	0.28	1.05	0.15
LHCb $Z \to ee$	1.07	0.19	1.03	0.08	1.08	0.22	1.04	0.11
CMS HM DY 8 TeV	0.98	-0.09	0.98	-0.08	0.99	-0.05	1.00	-0.02
CMS HM DY 13 TeV - combined channel	0.97	-0.14	0.97	-0.16	0.97	-0.14	0.97	-0.12
HL-LHC HM DY 14 TeV - neutral current - electron channel	1.01	0.03	1.03	0.08	1.04	0.10	1.21	0.53
HL-LHC HM DY 14 TeV - neutral current - muon channel	1.02	0.04	1.03	0.07	1.02	0.06	1.20	0.49
HL-LHC HM DY 14 TeV - charged current - electron channel	1.01	0.02	1.00	-0.00	1.15	0.42	2.97	5.56
HL-LHC HM DY 14 TeV - charged current - muon channel	0.97	-0.09	0.98	-0.07	1.11	0.31	2.75	4.94

Table D.2: Fit quality in fits contaminated with the W operator.

Appendix E

Contaminated PDF comparison

In Fig. E.1, we display the PDFs that are mostly affected by the new physics contamination in Scenario I, namely the anti-up and anti-down distributions at $Q = 2$ TeV in the large-x region. We see that for $\hat{Y} = 5 \cdot 10^{-5}$, PDFs are statistically equivalent to the baseline ones.

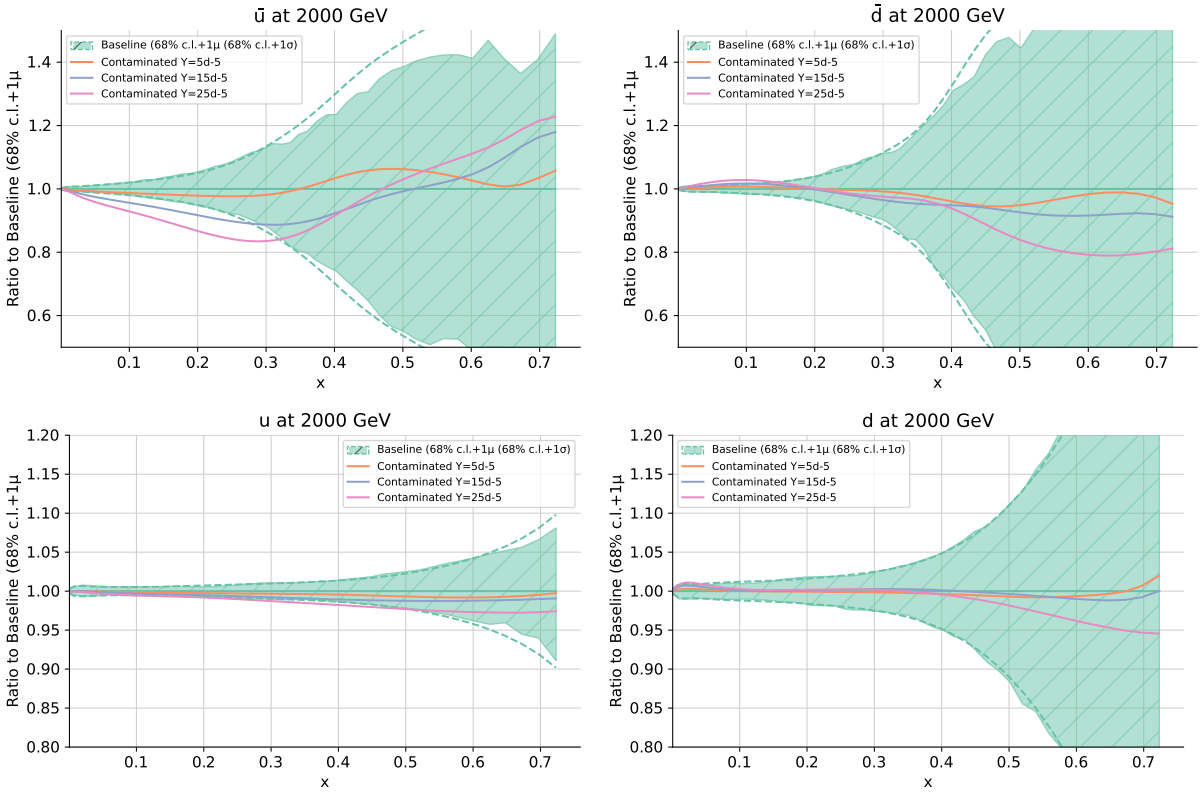


Figure E.1: Contaminated versus baseline anti-up (top-left panel), anti-down (top-right panel), up (bottom-left panel) and down (bottom-right panel) PDFs at $Q = 2$ TeV. The results are normalised to the baseline SM PDFs and the 68% C.L. band is displayed. Contaminated PDFs have been obtained by fitting the Monte Carlo pseudodata produced with $\hat{Y} = 5 \cdot 10^{-5}$ (orange line), $\hat{Y} = 15 \cdot 10^{-5}$ (blue line) and $\hat{Y} = 25 \cdot 10^{-5}$ (pink line) assuming the SM in the theory predictions.

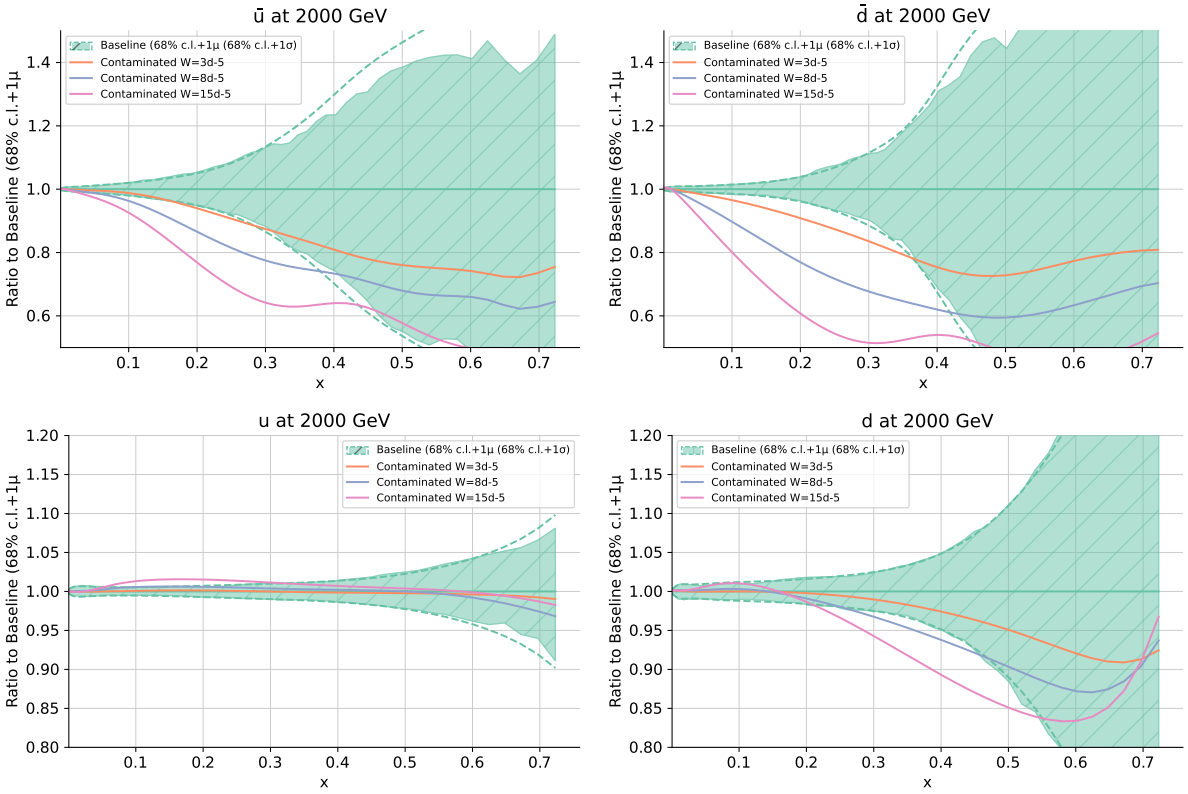


Figure E.2: Same as Fig. E.1 for $\hat{W} = 3 \cdot 10^{-5}$ (orange line), $\hat{W} = 8 \cdot 10^{-5}$ (blue line) and $\hat{W} = 15 \cdot 10^{-5}$ (pink line).

In Fig. E.2, we display the PDFs that are mostly affected by the Scenario II new physics contamination, namely the up, down, anti-up and anti-down distributions at $Q = 2$ TeV in the large-x region. We see that for the critical value $\hat{W} = 8 \cdot 10^{-5}$ the shift in the anti-quark PDFs is above the 2σ level for all of the distributions from $x \gtrsim 0.2$, apart from the up-quark PDF in which the shift is visible but below the 2σ level.